清華大學 Tsinghua University

基于自上而下图算法的语义图模型构建与应用

——以补充义副词为例

A Top-down Graph-based Tool for Modeling Classical Semantic Maps:
A Crosslinguistic Case Study of Supplementary Adverbs

刘柱 计算语言学 博士三年级

导师：刘颖教授

语义图模型

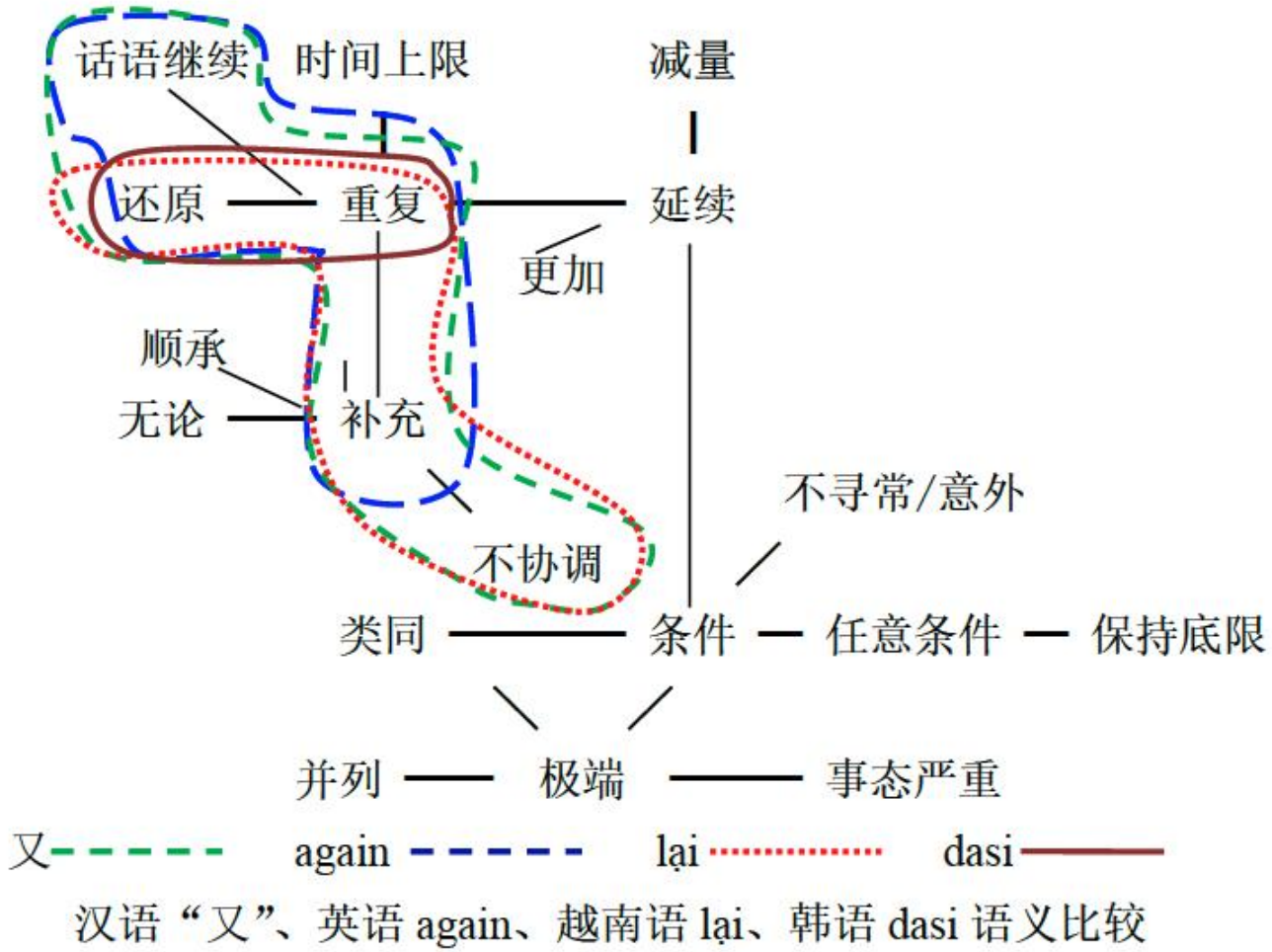
- 语义图模型（Semantic Map Models）使用图（graph）的形式展示了跨语言之间语义概念或者结构上的共性和规律
- 语言变异的有限性
  - 语义图模型构建的概念空间具有**结构**上的约束
  - 同一语言形式形成的语义图是**连通**的（连续性假说）
  - 不同语言形式构成的语义图尽可能存在蕴含共性
- 同形多义现象
  - 一词多义（多义词、共词化）
  - 一义多词（同义词）
- 虚词（多功能语法形式）

补充义副词

- 核心功能表示补充义的虚词
- 跨语言的多形式：还、又、也、在；also, too, again, still; も
- 多功能：补充、还原、重复等
  - 我明天还来。（重复）
  - 她买了菜，还做了饭。（补充）
  - If you fail your exam you will have to take it **again**. (重复)
  - After ten years in prison, he was a free man **again**. (还原)
- 受到学界的广泛关注(郭锐. 2010, Ying Zhang. 2017)

语义图模型+补充义副词

- （郭锐. 2010）收集了9种语言、共28种语言形式在18个功能下的数据，并且手工绘制了语义地图
- 然而手工绘制需要依次满足各个形式的连通性约束，当数据量变大时，过程会异常复杂



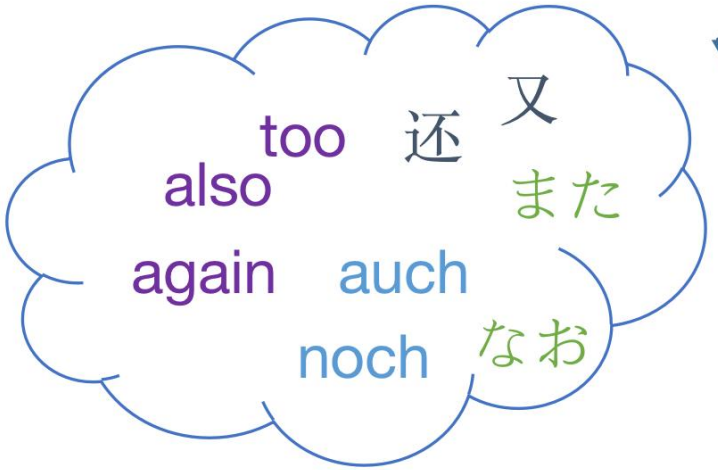
汉语“又”、英语 again、越南语 lại、韩语 dasi 语义比较

贡献

- 本文利用一种自上而下的图算法（工具）高效**自动构建**语义地图
- 设计不同的**评估指标**来挑选合适的图模型
- 以**补充义副词**为研究对象，得到专家标注类似的图模型，证明当前算法的有效性。

流程

补充义副词



多语言语料库

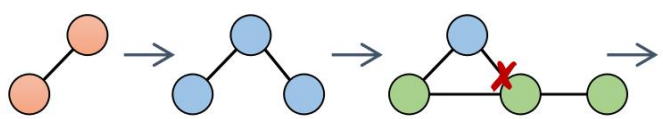
I want to see this movie **again**.
人死了还会活过来吗?
Bitte sag es **noch** einmal.
⋮

形式-功能 表格

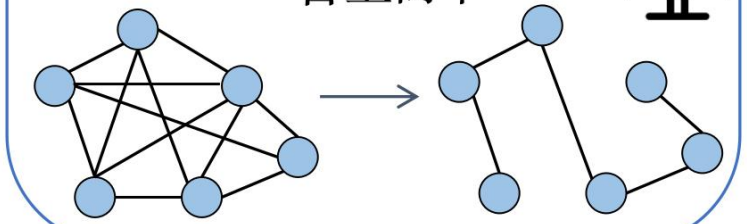
	补充	重复	条件	…
again	✓	✓	✗	
也	✓	✗	✗	
⋮				

语义图模型构建

自底而上



自上而下

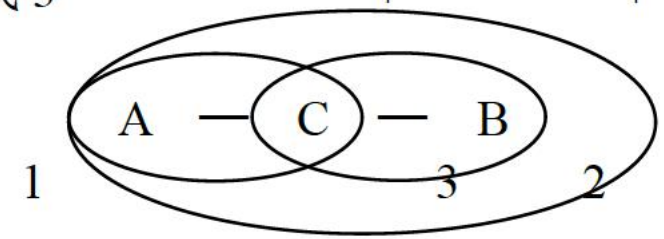


方法

（56）

	功能 A	功能 B	功能 C
形式 1	+	—	+
形式 2	+	+	+
形式 3	—	+	+

则：



非：

A — B — C

（形式 1 的功能无法连续分布）

- 语义图的结构
  - 点：功能/语义
  - 边：反映功能间的相似性

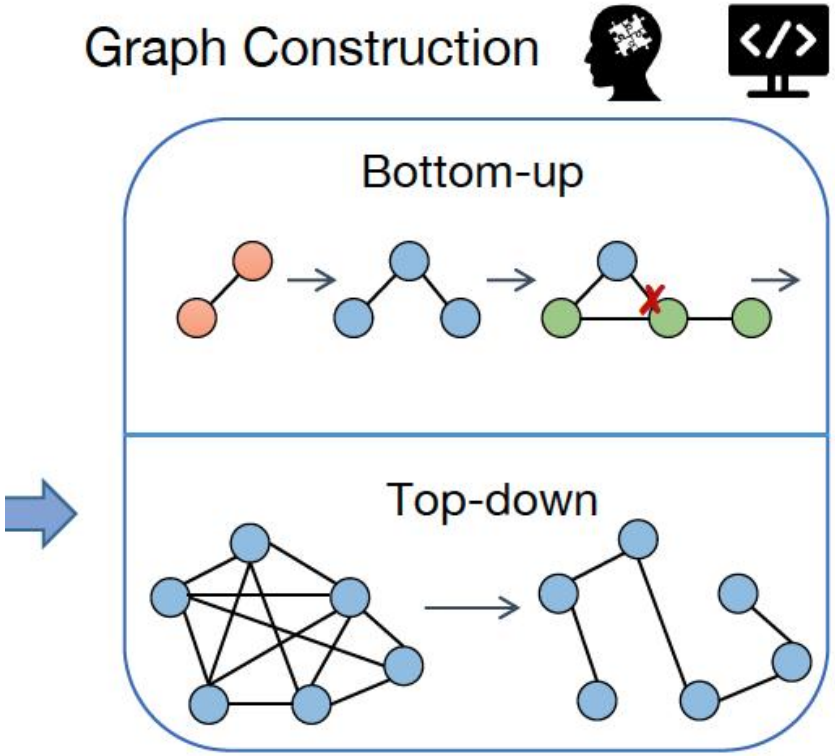
- 语义地图的**连续性假说H1**(Croft 2021)：特定语言形式对应的任何相关范畴都应映射在概念空间的一个连续区域（connected region）

方法

- 语义图的结构
  - 点：功能/语义
  - 边：反映功能间的相似性
  - 连通区域：某一语言形式拥有的所有功能

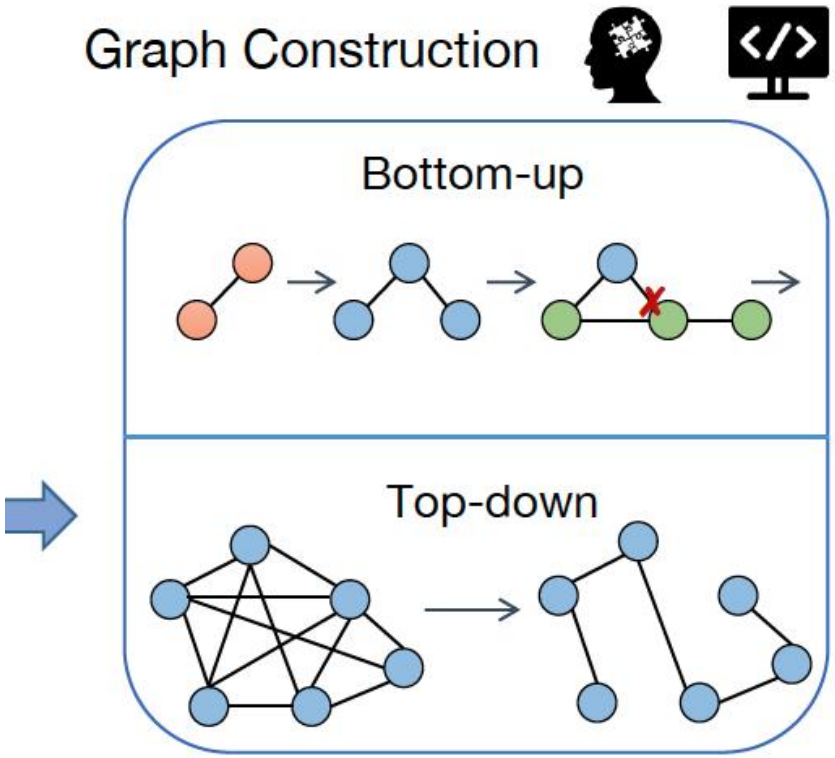
- 语义地图的**连续性假说H1**(Croft 2021)：特定语言形式对应的任何相关范畴都应映射在概念空间的一个连续区域（connected region）

- **自底向上**构建：逐形式地满足该条件
  - 缺点：数据量较大时候，复杂度提高；无法产生较多的候选图；无法自动化评估



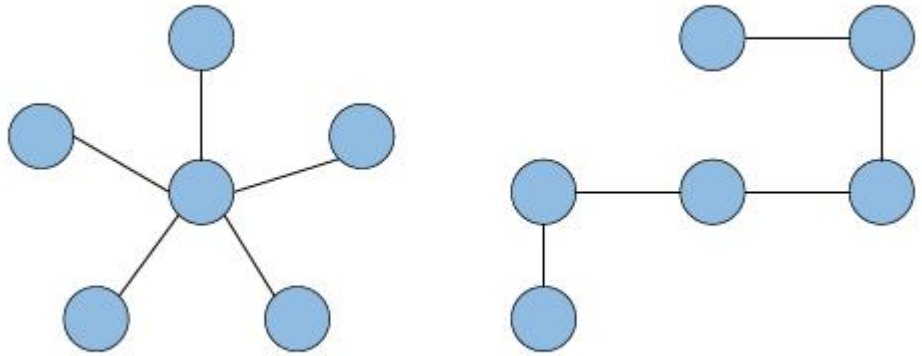
方法

- 自上而下构建考虑最终的网络满足：
  - 整体的连通性
    - 局部都是连通的
    - 如果存在“孤点”，说明该功能可以去掉
  - 无环
    - 环路会降低模型的预测力（Haspelmath, 2003），增加无用的边，从而降低模型的准确性
    - 现实为了满足覆盖率不得不加入环路
  - 权重最大
    - 边上的权重表示：两个功能共现（共词化）的次数
    - 反映了功能之间的相似性，应尽可能大
- 假设H2：最大生成树(maximum spanning tree)
  - 连通 + 无环 + 最大权重



方法 (挑选准则)

- 最大生成树算法生成很多候选子图
- 内部准则（表2）
  - 图的权重之和
  - 覆盖率
  - 精度
  - 度的方差(流式拓扑>星式拓扑)
- 外部准则
  - 与专家评估的准确性（逐元素对比）
  - 松下界：完全图
  - 紧下界：与GT不重叠的树



(a) std = 1.6　　(b) std = 0.5

Metric	Description	Trend
Size	Summed weights of edges	↑
Recall	Coverage rate of instances	↑
Precision	Accuracy of predicted instances	↑
Div_D	Standard deviation of degrees	↓
Acc	Matched rate compared to GT	↑

Table 2: Different metrics for evaluating the conceptual space. The trend shows the optimal direction for a better network.

实验

- 补充义副词（9种语言；28种语言形式；18个功能）（郭锐. 2010）

语言	词	类同	补充	重复	延续	更加	增加	减量	还原	条件	任意	极端	严重	无论	让步	上限	顺承	不协	意外	底限	续话	并列
汉语	还		补	重	延	更		减	还	条	任	极							意	底		
	又		补	重					还									不			续	
	也	类	补							条	任	极	严				顺			底		
	再		补	重	延	更			还					无		上						
藏语	ra	类	补							条		极	严									

其他语言包括：英语、德语、法语、俄语、日语、韩语、越南语

（1）根据该表格生成一个最初的带权图，权重表示**共现次数**

（2）最大生成图算法采用经典的克鲁斯卡尔算法，将图按照总权重大小进行排序

定量分析

- 覆盖率和精度的平衡
- 算出得出的最优图可以得到较大的覆盖率和很高的准确性，同时也保证可比的精度
- 无法保证100%的覆盖率，仍有四个语言形式无法满足，可能需要环的条件
  - 这些语言形式都有“条件”，它在人工构造中是环的中心点

Index	Size↑	Recall↑	Precision↑	Accuracy↑
C	286	1	0	50.0
LT	-	-	-	79.0
GT	91	1	0.20	1
0	90	85.7	0.17	92.6
1	89	82.1	0.21	91.4
2	89	82.1	0.44	90.1
3	88	82.1	0.34	91.4
4	88	78.6	0.50	88.9

Table 7: Evaluation of our generated graphs and baselines (denoted as complete graph C and ground truth GT). The index represents the first N maximum spanning trees, scaled by 10,000.

定量分析

- 测试一下度的标准差与准确率之间的关系
- 在随机的概念空间进行测试
  - 对于树而言，度本身就比较少
  - RG_1：完全随机的边
  - RG_2：边的出现与权重呈正比
- 结果可以得到一定的相关性

Round	RG_1	RG_2
1	-17.8	-22.1
2	-21.9	-22.4
3	-20.5	-19.2
4	-23.8	-21.7
5	-23.1	-24.1
Mean	-21.4	-21.9
Std. Dev.	2.13	1.58

Table 8: Pearson correlation between Div_D (diversity of degrees) and accuracy across five rounds. The mean and standard deviation for each round are also provided.

定性分析

- 一些关键节点(条件、重复补充延续)上重合性较高
- “条件”在原来形成了环
- 可以增加权重的分析
- 为专家提供一个初始版本，可以根据语言学的先验进行删改

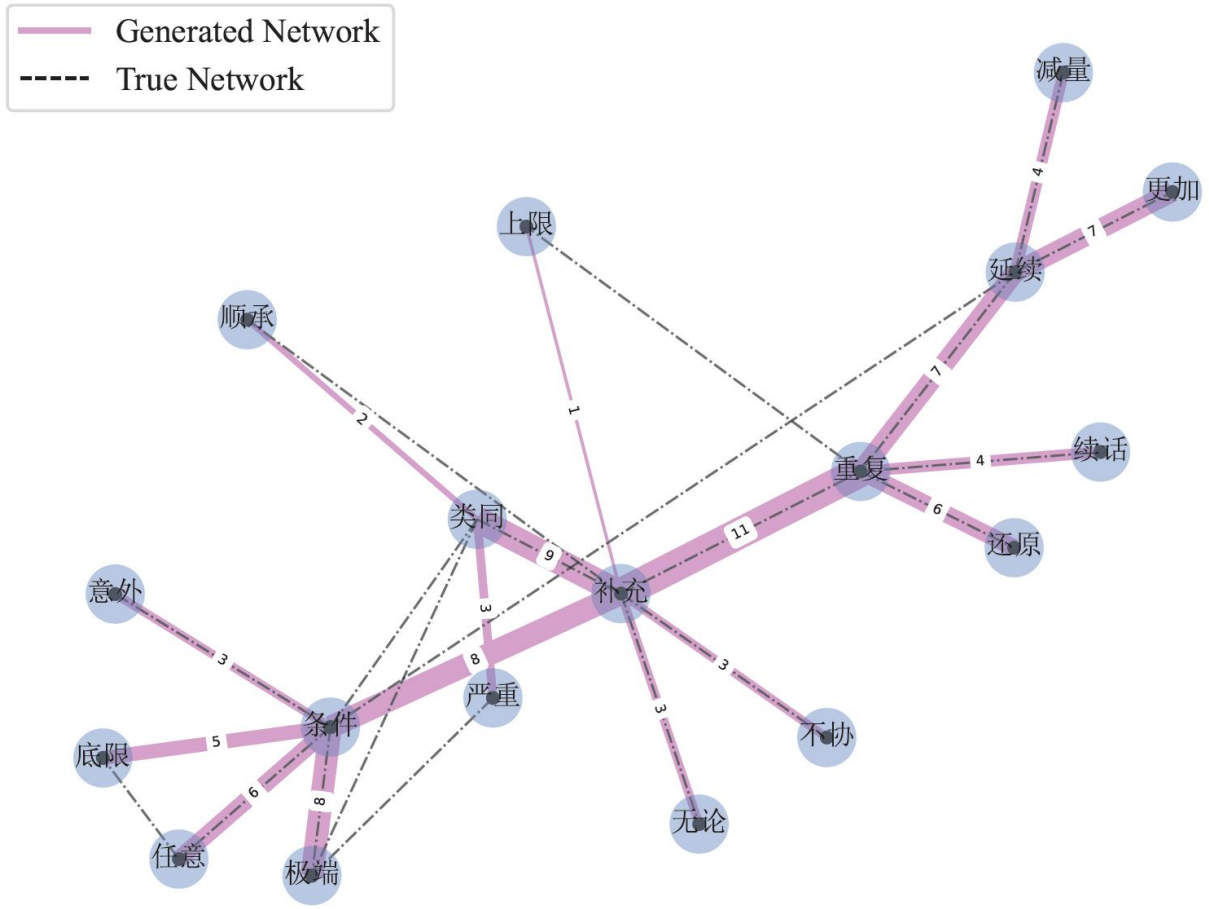


Figure 4: Tree of conceptual space with the largest size. The pink connections represent the network generated by our method, while the black dashed line indicates the ground truth as labeled by an expert. Numbers on the edge indicate the number of co-occurrences in a same word for the corresponding functions.

结论与展望

- 开发了一种自动化构建语义地图的算法和可视化工具
- 设计多个指标来评估语义地图

未来工作

- 利用大型语料库展开研究，例如使用在语料中共现的频率表示权重
- 利用语言模型进行研究，例如模型的中间表征来表示语义
- 词义选择的主观性和任意性，可以引入概率来刻画语义
- 融入时间信息

参考文献

- (郭锐. 2010) 副词的补充义与相关义项的语义地图. 中国语言的比较与类型学国际研讨会, Hong Kong, China.
- (Ying Zhang. 2017) Semantic map approach to universals of conceptual correlations: a study on multifunctional repetitive grams. Lingua Sinica, 3(1):7.
- (William Croft. 2001). Radical Construction Grammar:Syntactic Theory in Typological Perspective. Oxford University Press, Oxford.
- (Martin Haspelmath. 2003.) The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello, editor, The new psychology of language, volume 2, pages 211–243. Lawrence Erlbaum, Mahwah, NJ.

Q & A

https://github.com/RyanLiut/SemanticMapModel

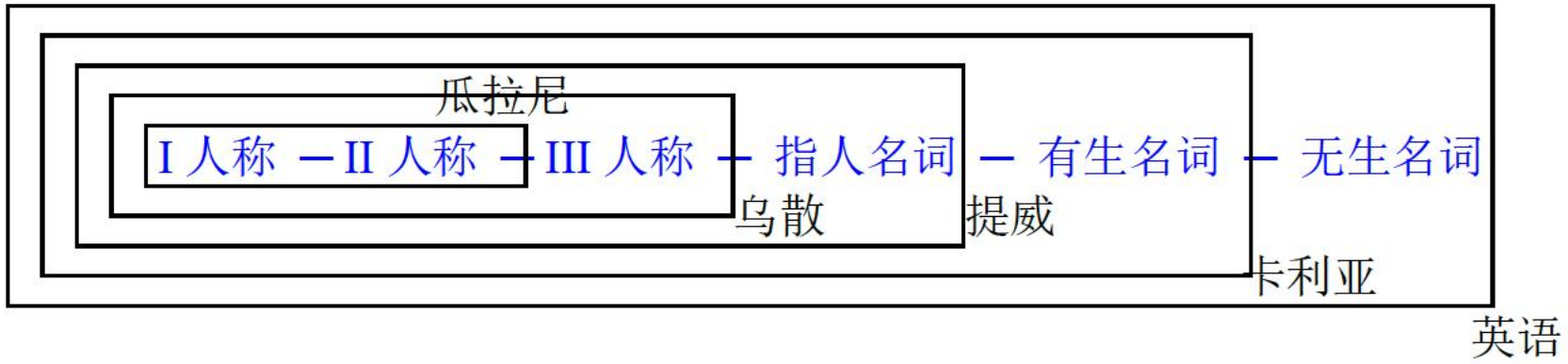
欢迎大家访问

Data

L	G	AF	SU	RE	CO	GD	DE	IS	CD	DC	PT	SC	WH	SE	SC	IC	UE	BL	DS
ZH	还	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	0
	又	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
	也	1	1	0	0	0	0	0	1	1	1	1	0	0	1	0	0	1	0
	在	0	1	1	1	1	0	1	0	0	0	0	1	1	0	0	0	0	0
BO	ra	1	1	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0
	tarong	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
EN	also	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	too	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	again	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	still	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	1	0
DE	auch	1	1	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0
	noch	0	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	0
FR	aussi	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	encore	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
RU	tbzhe	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	opyat’	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JA	も	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
	また	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	なお	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
KO	도	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
	더	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	또	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
	다시	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	아직	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
VI	cũng	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	0
	nữa	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	còn	0	0	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0
	lại	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0

Table 3: Form-function table for the Supplement-related semantic domain. Here, “L” represents languages and “G” denotes grams. Abbreviations for languages and functions are detailed in Tables 4 and 5. A value of “1” indicates that the gram corresponds to the function in at least one sentence.

无法满足的四个形式


瓜拉尼
I 人称
II 人称
III 人称
指人名词
有生名词
无生名词
乌散
提威
卡利亚
英语