清華大學 Tsinghua University

语义图模型

2025.03.13

刘柱

语义图模型

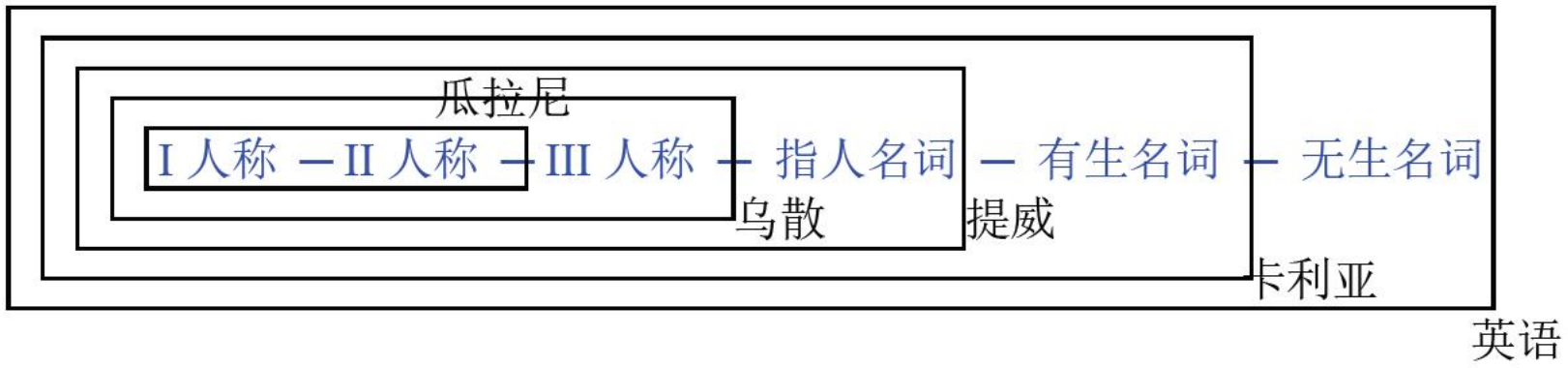
- 语义图模型（Semantic Map Models）使用图（graph/map）的形式展示了跨语言之间语义概念关联的共性和规律
  - 概念空间（Conceptual Space）以语义概念为结点，关联性强度为连边
- 提出的动机：语言变异的有限性
  - Greenberg关于语序类型学的研究之后，语言共性假说被广泛接受
  - 有限性：语言的差异不是任意的，而是受到一定条件约束的，有一些语言的变化方向是不可能的
  - 蕴含共性

语言变异有限性例子

- 名词是否有复数形式与名词的不同种类有关 *[Croft 2003]*
  - 瓦拉尼语：第一人称、第二人称
  - 乌散语：第一、第二、第三人称
  - 提威语：第一、二、三人称、指人名词
  - 卡利亚语：第一、二、三、指人、有生名词
  - 英语：第一、二、三、指人、有生、无生名词
  - 不存在语言：第一、第三 / 第一、无生名词
- 名词的生命度等级
  - 第一>第二>第三>有生>无生名词

语言变异有限性例子

- 可以根据上述材料绘制出概念空间
- 不可能语言出现在图中的非连通区域
- 概念空间制约着特定语言的语法形式的可能分布模式



Croft 2003

语义图模型/概念空间的构建

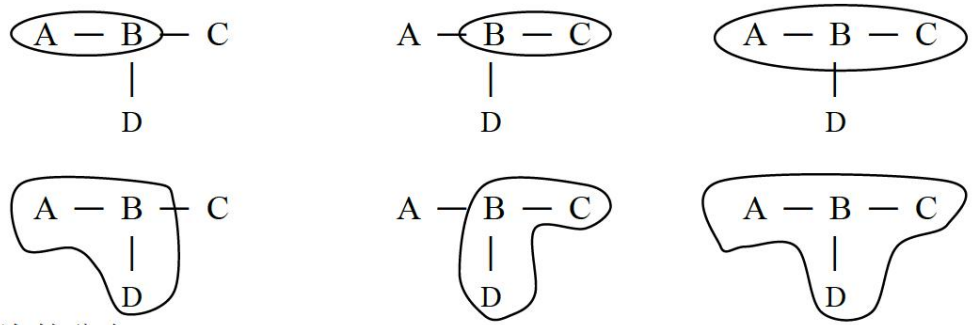
- 语言形式
  - 携带语义/功能的语言单元（能指）
  - 例如词、缀、结构/构式
- 概念
  - 语言形式的意义(meaning)
    - 虚词：语法功能(functions)，例如：让步、转折、提示、递进；有一些与语义角色相关（contextual meaning）；有一些是sense
    - 实词：词义（conventional meanings/sense），例如词典中列出来的
- 同形多义现象
  - 一个语言形式可以具有多个意义/功能
  - 实词：多义词；虚词：多功能语法形式（multifunctionality）
  - 由于虚词的语义更加不确定，语义图模型提供一个很好的表示工具

语义图模型/概念空间的构建

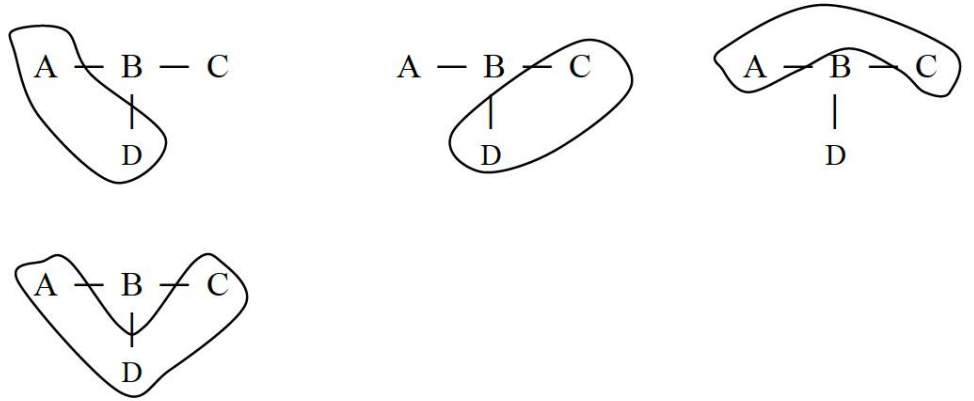
- 形式-功能对应关系
  - 一个语言形式具有的所有可能功能
  - 用一个0-1表格来表示
- 连通性假说 *[Croft 2003]*
  - 同一语言形式对应的功能所框定的语义子图是连通的
  - 出现在同一语言形式中的语义关联更加紧密
  - 共词化（colexification）
  - 例如：fly 具有“苍蝇”和“飞”两种语义，它们关联更加紧密
- 自下而上迭代构建
  - 逐案例满足（覆盖率）
  - 边不能太多；避免环出现（精确率/预测力）*[最小连接原则，Cysouw07]*

连通性假说

允许的分布：



不允许的分布：



	A	B	C	D
F1	1	1	0	0
F2	0	1	1	0
F3	1	1	1	0
F4	1	1	0	1
F5	0	1	1	1
F6	1	1	1	1

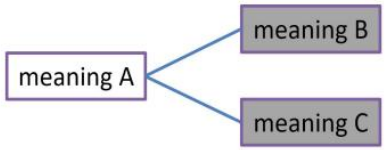


FIGURE 10a A simple semantic map

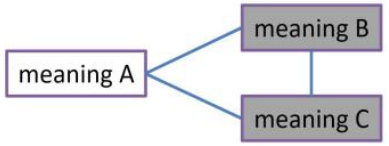
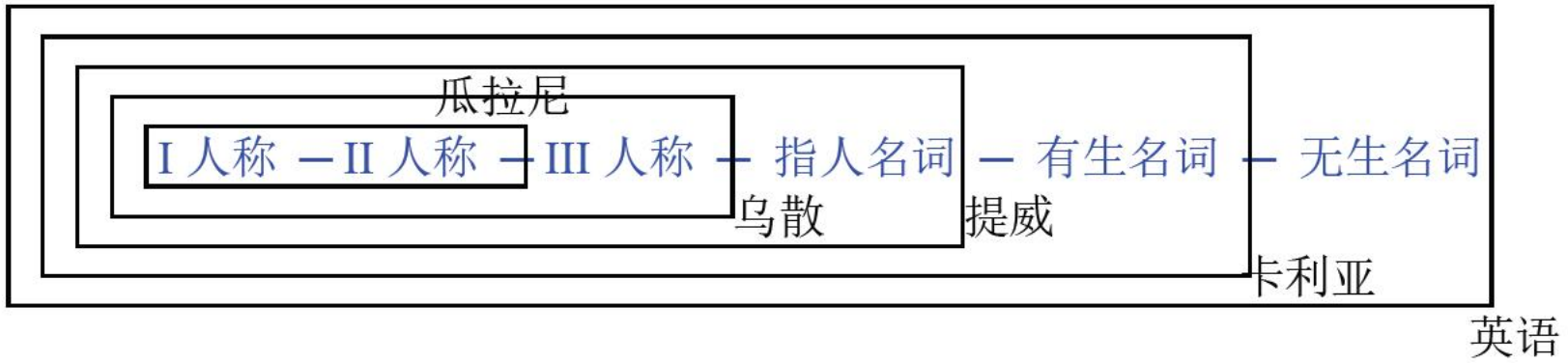


FIGURE 10b A vacuous semantic map

案例一

- 语言形式：复数标记
- 核心语义：复数概念



Croft 2003

案例二

- 语言形式：介词
- 核心语义：表示双及物结构中的接收者

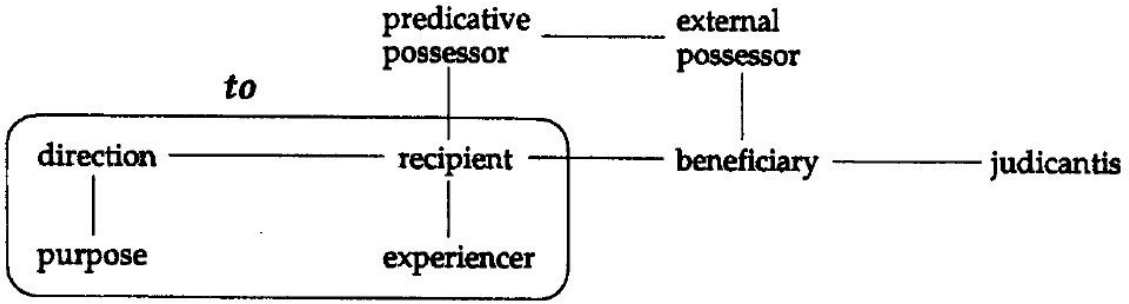


FIG. 8.1. A semantic map of typical dative functions/the boundaries of English *to*.

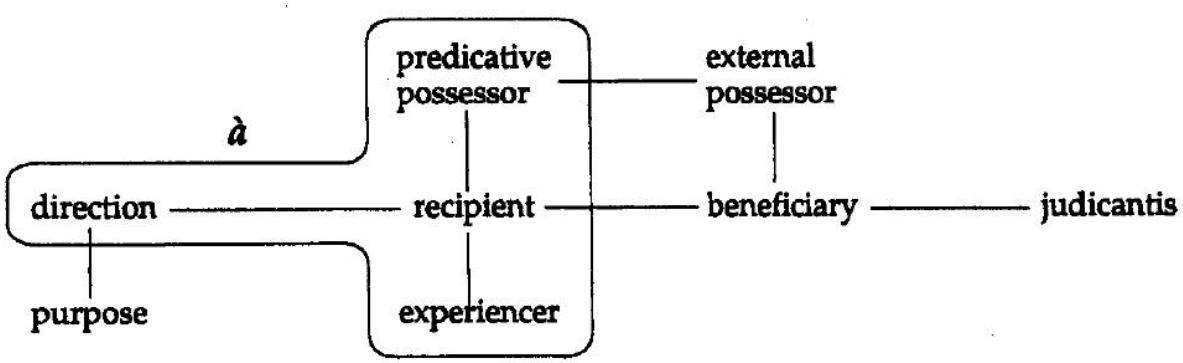
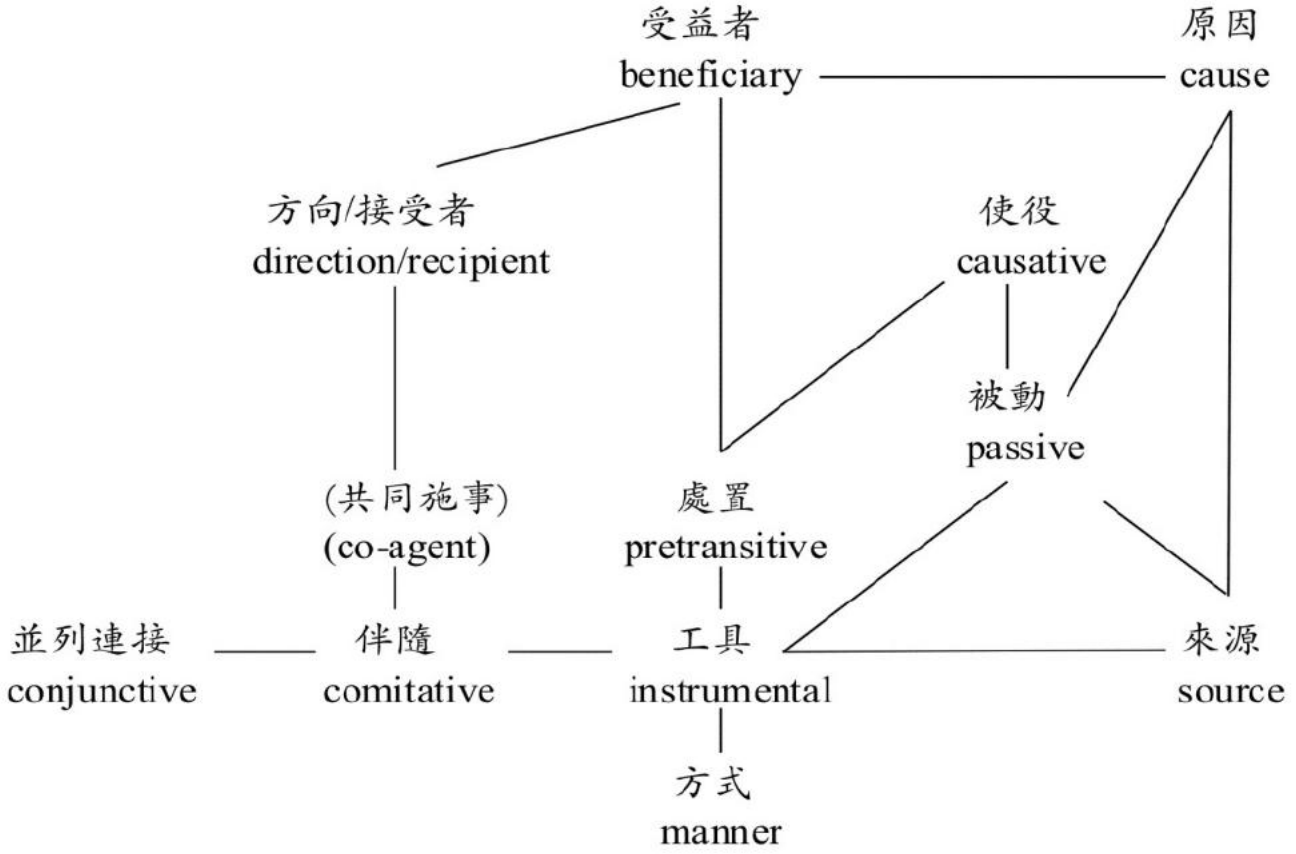


FIG. 8.2. The boundaries of French *à*.

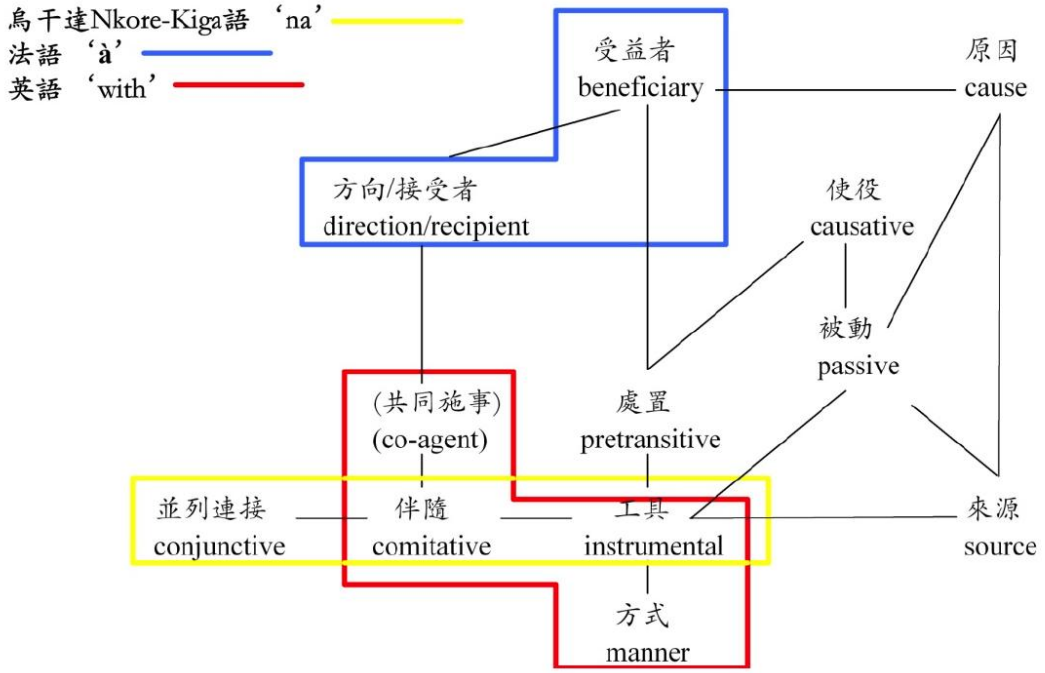
Haspelmath, M. 1999a

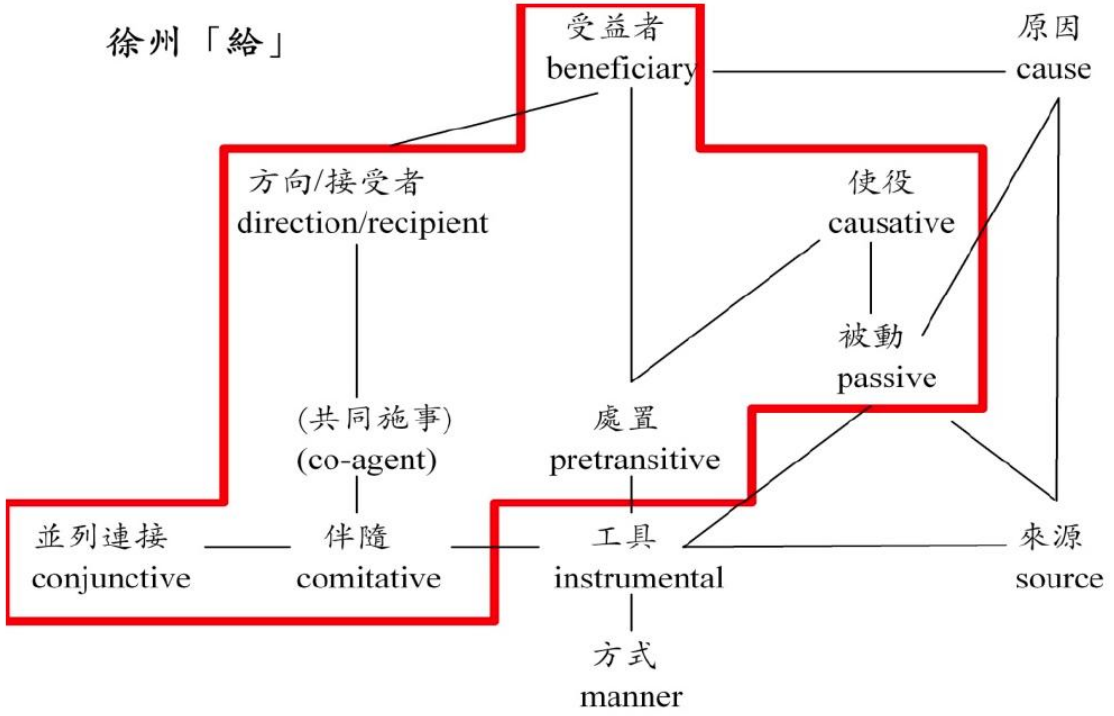
案例三

- 语言形式：介词
- 核心语义：处置义



张敏（2008）


烏干達Nkore-Kiga語 'na'
法語 'à'
英語 'with'
受益者
beneficiary
原因
cause
方向/接受者
direction/recipient
使役
causative
被動
passive
(共同施事)
(co-agent)
處置
pretransitive
並列連接
conjunctive
伴隨
comitative
工具
instrumental
來源
source
方式
manner

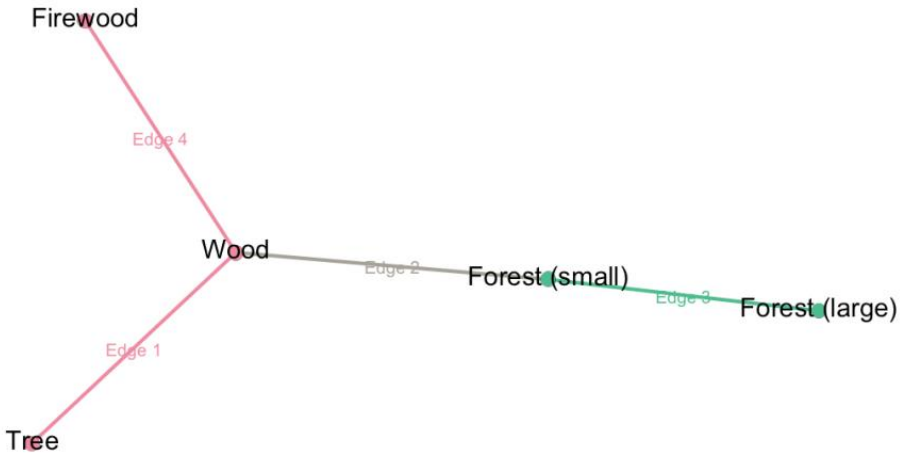

徐州「給」
受益者
beneficiary
原因
cause
方向/接受者
direction/recipient
使役
causative
被動
passive
(共同施事)
(co-agent)
處置
pretransitive
並列連接
conjunctive
伴隨
comitative
工具
instrumental
來源
source
方式
manner

案例四

- 语言形式：实词
- 核心语义：树木

TABLE 3 Lexical matrix for TREE/WOOD/FOREST in four languages

		MEANINGS				
		TREE	WOOD (mat.)	FIREWOOD	FOREST (small)	FOREST (large)
Danish	*Træ*	√	√	√	–	–
	Skov	–	–	–	√	√
French	*Arbre*	√	–	–	–	–
	Bois	–	√	√	√	(√)
	Forêt	–	–	–	(√)	√
German	*Baum*	√	–	–	–	–
	Holz	–	√	√	–	–
	Wald	–	–	–	√	√



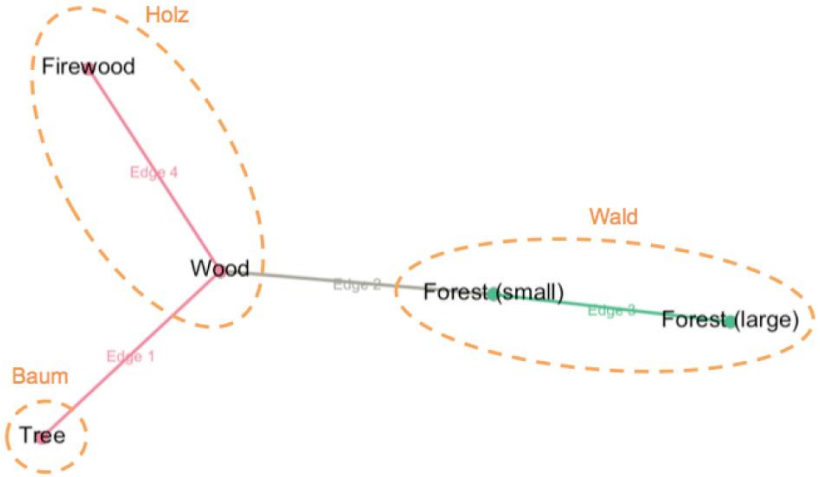


FIGURE 4 A semantic map inferred from the data of Table 2, with the German lexemes mapped onto the nodes

Polis 2018

案例五

- 语言形式：形容词
- 核心语义：属性义

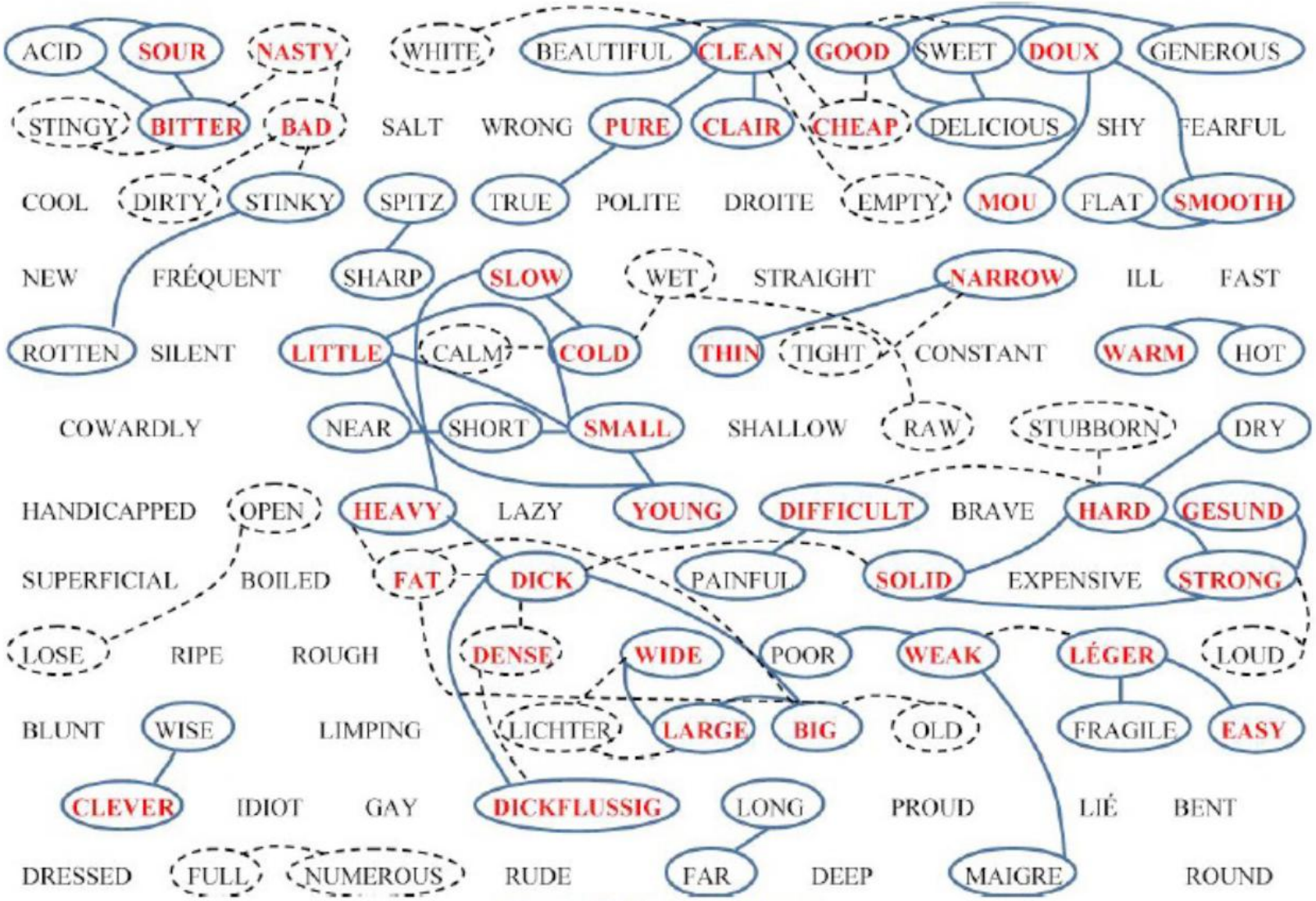


Figure 6: Conceptual map

Perrin 2010

案例六

- 语言形式：形容词
- 核心语义：SHARP义

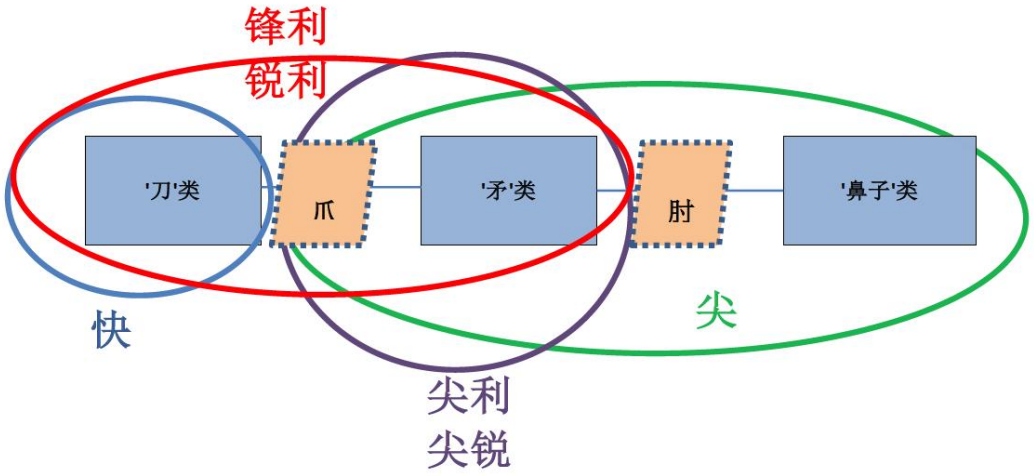


图 2.8　汉语 SHARP 语义场语义地图

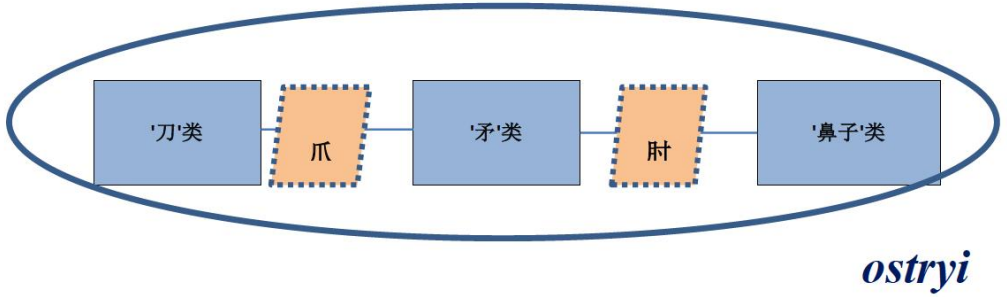


图 2.7　俄语 SHARP 语义场语义地图

俄语具体例子如下：

(10) ***ostryi***（锋利）*nozh*（刀）　　“**锋利**的刀”
(11) ***ostraya***（锋利）*igla*（针）　　“尖的针”
(12) ***ostryi***（锋利）*podborodok*（下巴）　　“尖下巴”

李亮 2015

参考文章

- Croft, William 2003 Typology and Universals(2nd edition). Cambridge: Cambridge Universitiy Press.
- Cysouw M. (2007a) Building semantic maps: The case of person marking. In: Wälchli B., Miestamo M. (eds) New challenges in typology. Mouton de Gruyter, Berlin, pp 225–248
- Haspelmath M. External possession in a European areal perspective[J]. Typological studies in language, 1999, 39: 109-136.
- 张 敏（2008b）“汉语方言处置式标记的类型学地位及其它”，北京大学中国语言学研究中心演讲稿，2008年1月8日。
- Georgakopoulos T, Polis S. The semantic map model: State of the art and future avenues for linguistic research[J]. Language and Linguistics Compass, 2018, 12(2): e12270.
- Loïc-Michel Perrin. 2010. Polysemous qualities and universal networks, invariance and diversity. Linguistic Discovery, 8:1–22.
- 李亮《词汇类型学视角的汉语物理属性形容词研究》（北京大学博士论文，2015）

讨论环节

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

背景

- 经典语义图模型由语言学专家自下而上构造
  - 自下而上：依次考虑每个词形对应的功能，边的构建过程是由少到多
  - 优势：构造图的过程中，可以产生一些语言学的分析

节点的排列有时需要结合语义分析。比如：

（58）		补充	类同	顺承	
汉语	“也”	+	+	+	补充：他会骑马，也会开车
德语	auch	+	+	+	类同：他会骑马，我也会骑马。
英语	also	+	+	—	顺承：50 年过去了，我也从黑发到灰发，直到现在的白发。

这三个功能的两种可能的排列：1. 补充——类同——顺承

2. 类同——补充——顺承

也就是“顺承”是与“补充”直接关联，还是与“类同”直接关联，仅仅根据语言事实无法确定，需对三个功能作语义分析。由于“顺承”可分析为“补充”的语境义吸纳，所以选择 2 较好。

可见，概念空间的构建主要是根据语言事实归纳的结果。

背景

- 经典语义图模型由语言学专家自下而上构造
  - 自下而上：依次考虑每个词形对应的功能，边的构建过程是由少到多
  - 优势：构造图的过程中，可以产生一些语言学的分析
  - 不足
    - 数据规模（词形、功能等）更大，人工选边的选择太多、且很容易产生多余的边/环，效率很低
    - 边的选择相对主观、边没有权重等反映重要性
    - 语义学的分析可以在图构建结束后发生，构建时候根据语言事实满足覆盖率应该是首要条件
  - 相关工作：以往人工构造实例；程序算法*[马腾]*
- 第二代语义图模型全局构造
  - 自上而下：构造出所有功能结点都具有连边的完全图
  - 连边权重：功能共现的次数

第二代语义图-例

Cysouw(2003)把人称标记（包括人称代词、人称前后缀和动词的人称变化形式等）的功能分析为 8 个基元（primitives）。

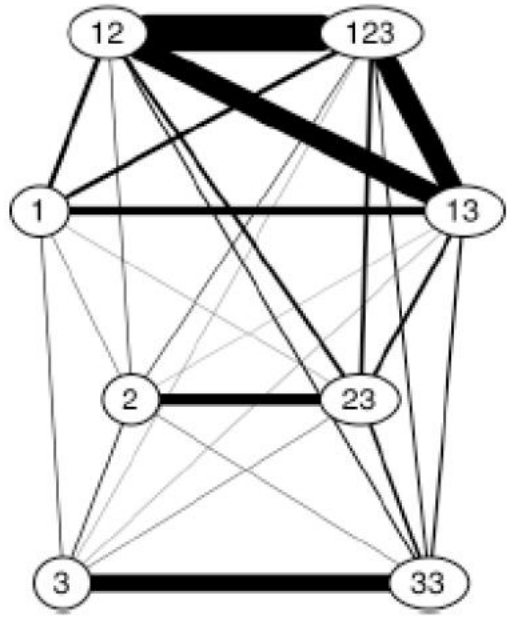
序号	基元	指称意义	英语
1	1	说话人	I、我
2	2	听话人	you、你
3	3	除 1 和 2 外的其他人	he/she/it、他
4	12	说话人和听话人（双数包括式）	we、咱们/我们
5	123	说话人、听话人和其他人（复数包括式）	we、咱们/我们
6	13	说话人和第三人（排除式）	we、我们
7	23	听话人和第三人	you、你们
8	33	多个第三人（不包括说话人和听话人）	they、他们

表 2：八个人称标记基元（Cysouw 2003）　（8 个基元中无 11、22）

Cysouw （2003）

范畴(基元结合体)	大致意思	频率	范畴(基元结合体)	频率
3/33	第三人称	125	1/2	3
12/123/13	第一人称复数(我们)	100	1/2/3	3
12/123	包括式(咱们)	97	12/13	2
2/23	第二人称	84	13/23	2
1/12/123/13	第一人称	35	3/23	2
1/13	排除式	29	12/123/23	2
12/123/13/23	非第三人称复数	18	1/12/123/13/23	2
23/33	非第一人称复数	17	123/13/23	1
12/123/13/33	非第二人称复数	11	13/33	1
1/3	非第二人称单数	10	1/12	1
2/3	非第一人称单数	7	1/23	1
2/3/23/33	非第一人称复数	6	12/123/33	1
3/13/33		5	1/12/123	1
2/12/123/13		5	3/12/123/33	1
12/123/13/23/33		5	1/2/12/123/13/23	1
2/13/23		4	2/12/123/13/23/33	1
2/12/123/23		4	1/2/12/123/13/23/33	1
123/13		3		591

表 3：35 种基元组合体频率数据表（Cysouw 2003）



背景

- 第二代语义图
  - 缺点：连边太多、所有可能出现的语言形式过多（导致准确率下降）；存在很多冗余的连边
  - 完全连通图是满足连通性的平凡解
- 如何删边？
  - 删除尽可能多的边，仍然要保持连通性假设
  - 手工分析*[郭锐讲稿]*
  - 设置阈值*[陈振宇，2015]*

贡献

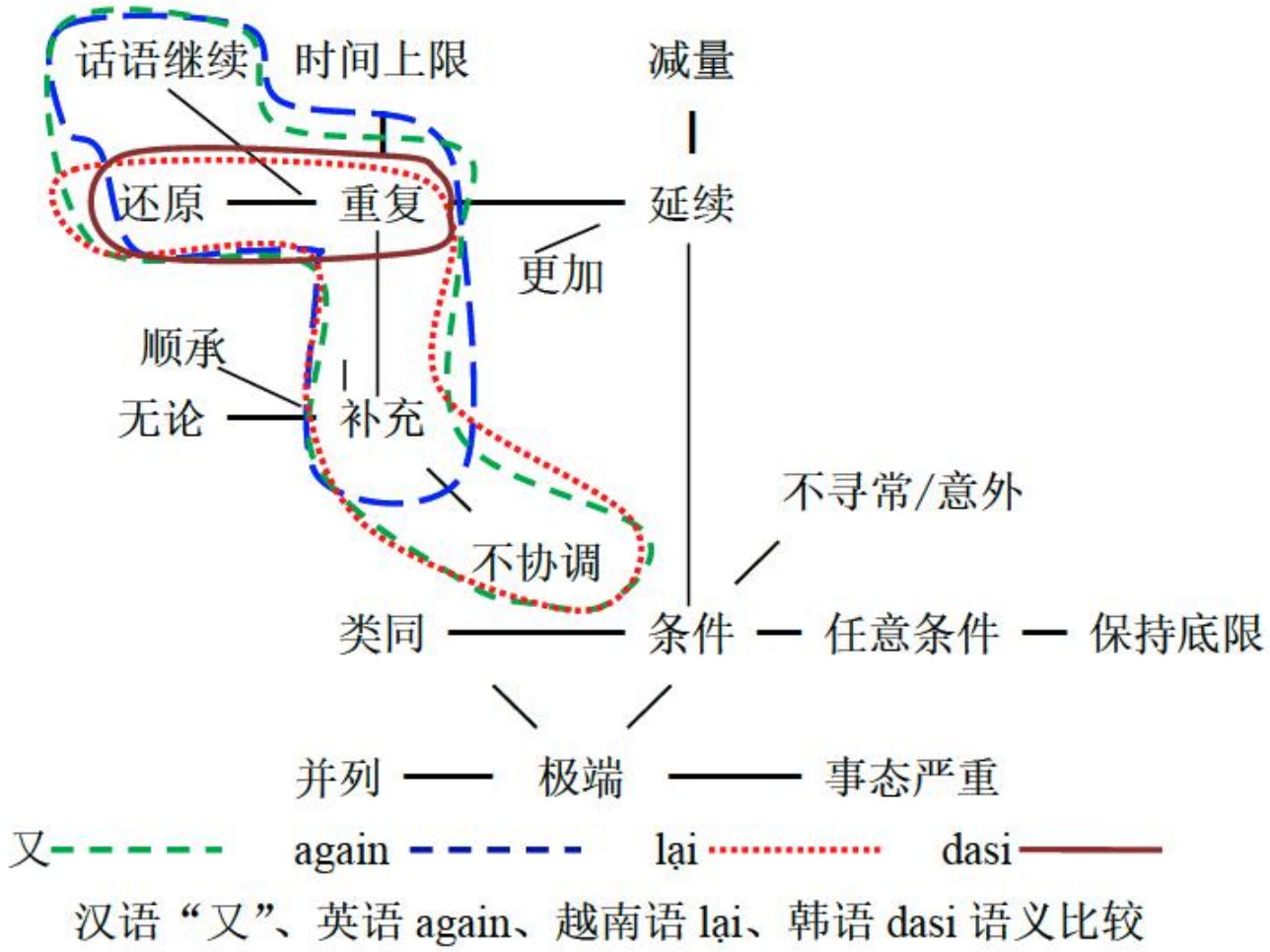
- 本文利用一种自上而下的图算法高效**自动构建**语义地图
- 设计不同的**评估指标**来挑选合适的图模型
- 以**补充义副词**为研究对象，得到专家标注类似的图模型，证明当前算法的有效性。

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

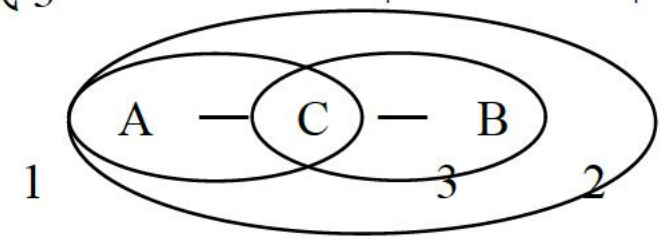
流程


补充义副词
too
还
又
also
また
again
auch
noch
なお
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

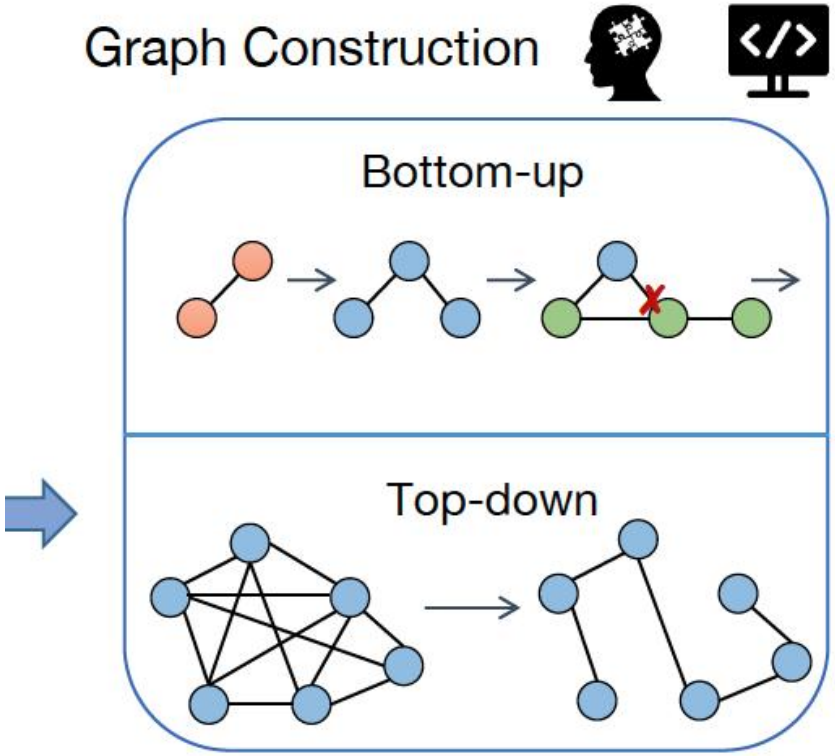
- 语义地图的**连续性假说H1**(Croft 2021)：特定语言形式对应的任何相关范畴都应映射在概念空间的一个连续区域（connected region）

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

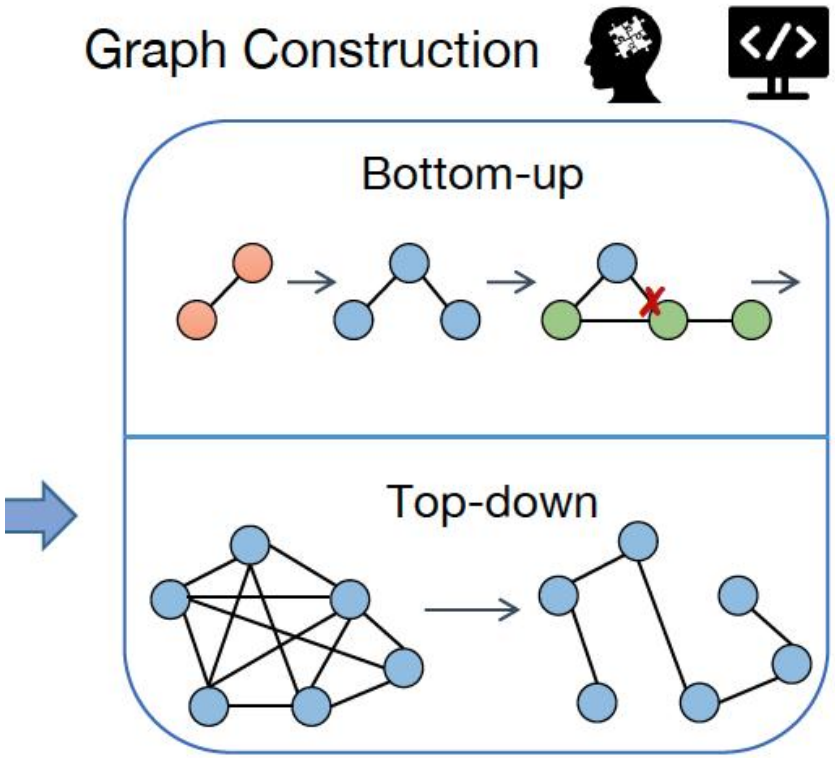
- 语义地图的**连续性假说H1**(Croft 2021)：特定语言形式对应的任何相关范畴都应映射在概念空间的一个连续区域（connected region）

- **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



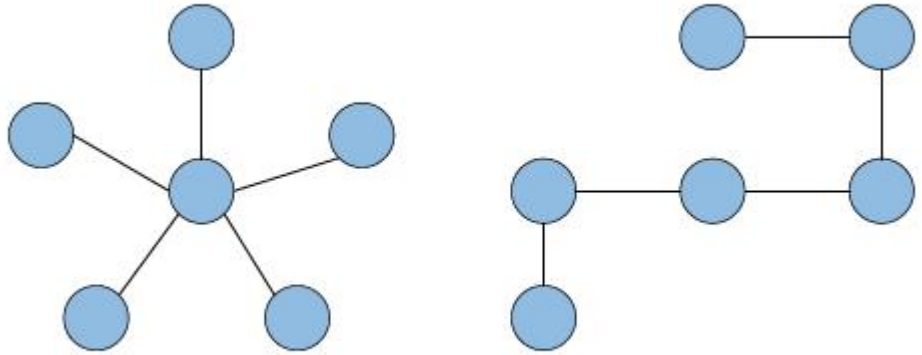
方法

- 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

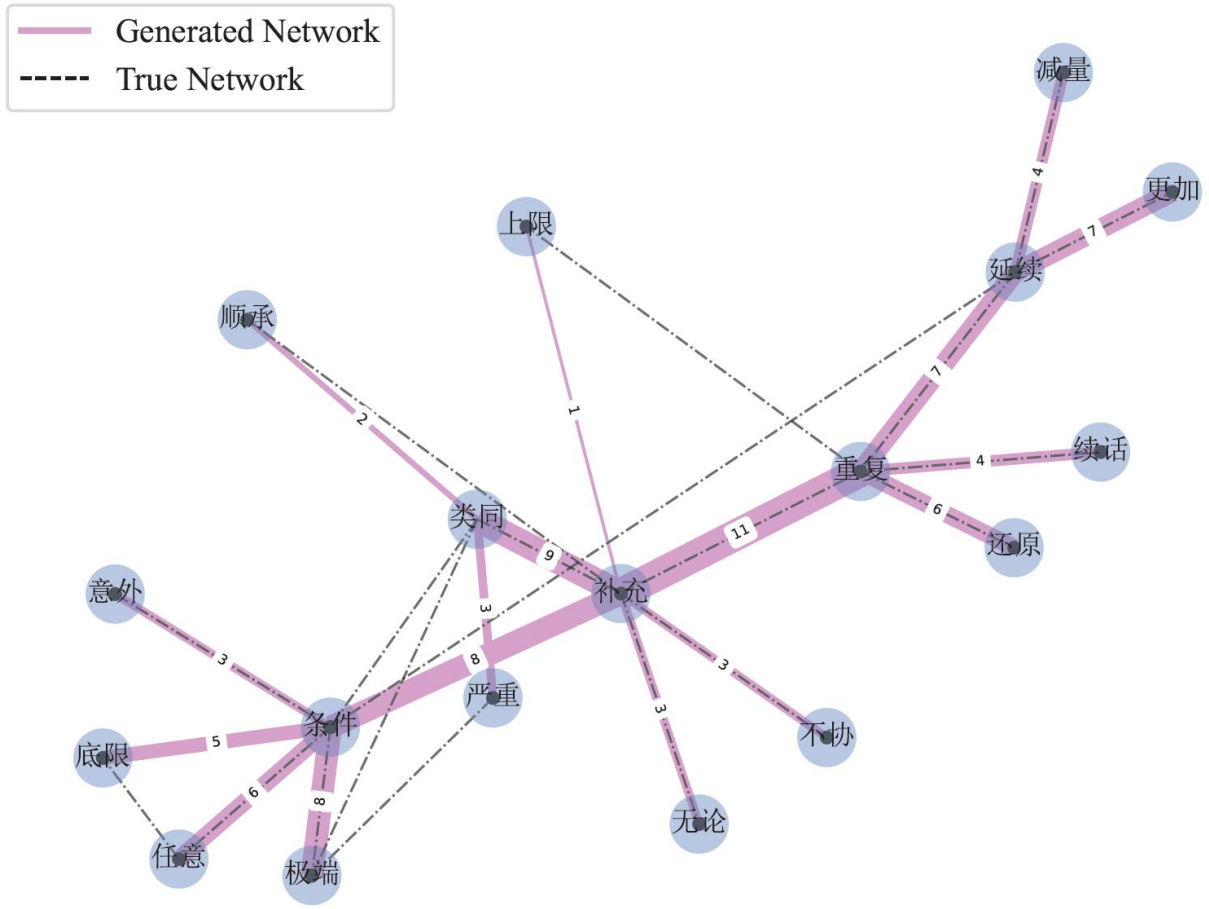


Figure 4: Tree of conceptual space with the largest size. The pink connections represent the network generated by our method, while the black dashed line indicates the ground truth as labeled by an expert. Numbers on the edge indicate the number of co-occurrences in a same word for the corresponding 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.
- Cysouw, Michael(2007a). Building Semantic Maps: the Case of Person Marking. In: Matti Miestamo & Bernhard Wälchli (eds)., New Challenges in typology: Broadening the horizons and redefining the foundations. Berlin: Mouton, P225-248.
- 陈振宇, 陈振宁. 通过地图分析揭示语法学中的隐性规律——“加权最少边地图”[J]. 中国语文, 2015 (5): 428-438.

Q & A

(Code)

(Paper)

https://github.com/RyanLiut/SemanticMapModel

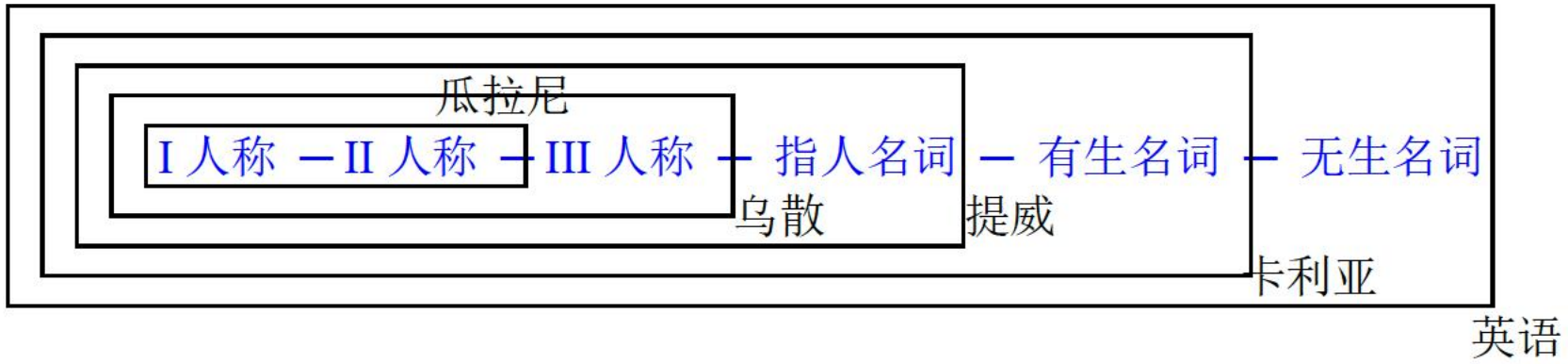
欢迎大家访问

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