

# CQL 的實際應用

## Grammatical Collocations

HOCOR 2020

廖永賦  
台大語言所

# Why Build Corpora with Corpus Engines?

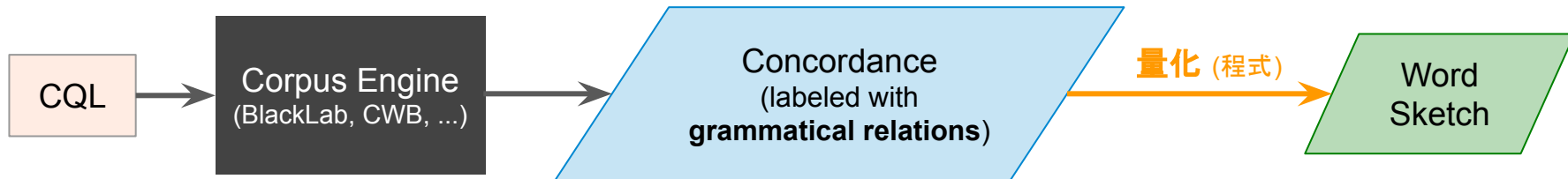
- Tools: [BlackLab](#), [\(No\)Sketch Engine](#), [CWB](#)
- 預建功能 (Concordance, Collocation, Word Frequency Lists, ...)
- 搜尋速度快 (建立索引)
- Query Language (e.g., **CQL**)

# 除了找 Concordance, CQL 還可以？

- 找出 concordance 後, (除了人工檢視外) 還能有許多應用...
- 中研院[中文詞彙特性速描系統 \(Chinese Word Sketch\)](#) (Huang et al., 2005)

喝	21	14	7	0	-7	-14	-21	吃									
SentObject_of 1287 4841 4.9 7.5					Modifier 5396 16648 3.6 4.5					Object 19799 40340 4.5 3.7							
喜歡		<a href="#">215</a>	<a href="#">750</a>	59.7	72.8	少		<a href="#">114</a>	<a href="#">493</a>	47.0	64.4	酒		<a href="#">6001</a>	<a href="#">18</a>	108.7	9.1
試		<a href="#">34</a>	<a href="#">465</a>	34.9	71.7	多		<a href="#">362</a>	<a href="#">1567</a>	47.5	62.8	牛奶		<a href="#">440</a>	<a href="#">24</a>	66.7	18.5
愛		<a href="#">236</a>	<a href="#">798</a>	58.0	69.9	一起		<a href="#">188</a>	<a href="#">412</a>	42.8	46.2	東西		<a href="#">22</a>	<a href="#">830</a>	12.6	55.3
嗜		<a href="#">11</a>	<a href="#">89</a>	29.5	58.4	不		<a href="#">1037</a>	<a href="#">2191</a>	45.3	45.9	奶		<a href="#">138</a>	<a href="#">183</a>	48.3	47.8
拒		<a href="#">13</a>	<a href="#">186</a>	23.3	56.3	常		<a href="#">116</a>	<a href="#">238</a>	41.5	44.5	稀飯		<a href="#">24</a>	<a href="#">50</a>	34.0	41.6
顧不上		<a href="#">12</a>	<a href="#">69</a>	31.5	54.9	天天		<a href="#">35</a>	<a href="#">92</a>	35.5	43.9	碗		<a href="#">26</a>	<a href="#">100</a>	23.8	37.6
敢		<a href="#">44</a>	<a href="#">227</a>	33.6	50.4	沒		<a href="#">93</a>	<a href="#">352</a>	31.7	42.7	水		<a href="#">381</a>	<a href="#">16</a>	37.3	1.9
請		<a href="#">53</a>	<a href="#">243</a>	27.7	39.8	邊		<a href="#">43</a>	<a href="#">176</a>	28.5	41.2	習慣		<a href="#">194</a>	<a href="#">209</a>	35.8	31.4
喜愛		<a href="#">27</a>	<a href="#">51</a>	31.7	33.6	連		<a href="#">24</a>	<a href="#">205</a>	19.3	39.8	奶水		<a href="#">23</a>	<a href="#">19</a>	35.5	29.6
怕		<a href="#">10</a>	<a href="#">58</a>	18.0	33.1	只		<a href="#">163</a>	<a href="#">454</a>	32.9	38.6	母奶		<a href="#">23</a>	<a href="#">26</a>	34.3	32.7
涉嫌		<a href="#">45</a>	<a href="#">11</a>	28.7	8.8	不要		<a href="#">125</a>	<a href="#">259</a>	35.7	37.9	母乳		<a href="#">39</a>	<a href="#">43</a>	31.1	28.5
知道		<a href="#">30</a>	<a href="#">74</a>	24.2	27.8	給他		<a href="#">16</a>	<a href="#">71</a>	23.4	37.6	酸奶		<a href="#">20</a>	<a href="#">10</a>	31.0	19.5

# Chinese Word Sketch 的概念 (以 Modifier + V 為例)



**Mod:** [pos="D.\*" & pos!="DE"] [word="地"]? **V:** [pos="V.\*" & pos!="VH.\*"] [pos!="DE"]

逐漸D

地DE

跟上VC

腳步Na

稍微D

擠出來VB

，COMMACATEGORY

很Dfa

愛VL

笑VA

<sup>1</sup> CKIP tag set <https://github.com/ckiplab/ckiptagger/wiki/POS-Tags>

<sup>2</sup> PTT 語料庫 <http://140.112.147.132:9898/concordance>

<sup>3</sup> CQL 範例 **Mod:** [pos="D.\*" & pos!="DE"] [word="地"]? **V:** [pos="V.\*" & pos!="VH.\*"] [pos!="DE"]

# BlackLab API

```
{
  "summary": {...}, // 17 items
  "hits": [
    {
      "docPid": "M.1546281742.A.844",
      "start": 10,
      "end": 12,
      "captureGroups": [...], // 2 items
      "left": {
        "punct": [...], // 3 items
        "pos": [...], // 3 items
        "word": [...] // 3 items
      },
      "match": {
        "punct": [...], // 2 items
        "pos": [...], // 2 items
        "word": [...] // 2 items
      },
      "right": {
        "punct": [...], // 3 items
        "pos": [...], // 3 items
        "word": [...] // 3 items
      }
    },
    {...}, // 7 items
    {...} // 7 items
  ],
  "docInfos": {...} // 2 items
}
```

hits: concordance

captureGroups

left: context

match: keywords matching CQL

right: context

```
"captureGroups": [
  {
    "name": "Mod",
    "start": 10,
    "end": 11
  },
  {
    "name": "V",
    "start": 11,
    "end": 12
  }
],
```

CQL Label name

Position in corpus

BlackLab API 回傳 JSON 格式: [bit.ly/blacklap-api](https://bit.ly/blacklap-api)

API 搜尋參數: <https://inl.github.io/BlackLab/blacklab-server-overview>

CQL: **Mod**: [pos="D.\*" & pos!="DE"] [word="地"]? **V**: [pos="V.\*" & pos!="VH.\*"] [pos!="DE"]

# 程式手作: Collostructional Analysis

# Collostructional Analysis

- **Collexeme analysis** (Stefanowitsch & Gries, 2003)
  - 衡量句式與其 lexical slot 內的詞彙的共現傾向  
e.g., 「把」字句中之**動詞**使用偏好
- **Distinctive collexeme analysis** (Gries & Stefanowitsch, 2004)
  - 比較兩種 (or 多種) 句式中, 相應位置之 lexical slot 的偏好  
e.g., 「把」字句 vs. 「將」字句, 句中之**動詞**使用偏好
- **Co-varying collexeme analysis** (Stefanowitsch & Gries, 2005)
  - 衡量同一句式下的兩個 lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的**賓語**與**動作**, 如: 把 **時間**(slot1) **花**(slot2) 在...

	$L_j$	$\neg L_j$
<b>C</b>	<i>a</i>	<i>b</i>
$\neg \mathbf{C}$	<i>c</i>	<i>d</i>

	$L_j$	$\neg L_j$
<b>C<sub>1</sub></b>	<i>a</i>	<i>b</i>
<b>C<sub>2</sub></b>	<i>c</i>	<i>d</i>

	$L_{\text{Slot 1}}$	$\neg L_{\text{Slot 1}}$
$L_{\text{Slot 2}}$	<i>a</i>	<i>b</i>
$\neg L_{\text{Slot 2}}$	<i>c</i>	<i>d</i>

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)
  - 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	¬重心
置		
¬置		



# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向

e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	$L_{\text{Slot 1}}$	$\neg L_{\text{Slot 1}}$
$L_{\text{Slot 2}}$	<i>a</i>	<i>b</i>
$\neg L_{\text{Slot 2}}$	<i>c</i>	<i>d</i>

text

你要把 重心 置 於家人身上，  
而不是把 重心 放 在工作。  
整天把 心思 放 在工作卻不管家人實在很糟  
應該多把 時間 花 在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	$\neg$ 重心
置	1	
$\neg$ 置		

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	¬重心
置	1	
¬置		

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	¬重心
置	1	
¬置		

1

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	¬重心
置	1	0
¬置		

1

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	¬重心
置	1	0
¬置	1	

1

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如:把 時間(slot1) 花(slot2) 在...

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	¬重心
置	1	0
¬置	1	

1

2

4

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	¬重心	
置	1	0	1
¬置	1		3
	2		4

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)
  - 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如:把 時間(slot1) 花(slot2) 在...

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	¬重心	
置	1	0	1
¬置	1		3
	2	2	4



# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	$L_{\text{Slot 1}}$	$\neg L_{\text{Slot 1}}$
$L_{\text{Slot 2}}$	<i>a</i>	<i>b</i>
$\neg L_{\text{Slot 2}}$	<i>c</i>	<i>d</i>

text

你要把重心置於家人身上，  
而不是把重心放在工作。  
整天把心思放在工作卻不管家人實在很糟  
應該多把時間花在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	重心	$\neg$ 重心	
置	1	0	1
$\neg$ 置	1	2	3
	2	2	4

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)
  - 衡量同一句式下的兩個lexical slots 內的詞彙的共現傾向  
e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

	$L_{\text{Slot 1}}$	$\neg L_{\text{Slot 1}}$
$L_{\text{Slot 2}}$	<i>a</i>	<i>b</i>
$\neg L_{\text{Slot 2}}$	<i>c</i>	<i>d</i>

text

你要把 重心 置 於家人身上，  
而不是把 重心 放 在工作。  
整天把 心思 放 在工作卻不管家人實在很糟  
應該多把 時間 花 在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	心思	$\neg$ 心思
放	1	1
$\neg$ 放	0	2

	重心	$\neg$ 重心
放	1	1
$\neg$ 放	1	1

	重心	$\neg$ 重心
置	1	0
$\neg$ 置	1	2

	時間	$\neg$ 時間
花	1	0
$\neg$ 花	0	3

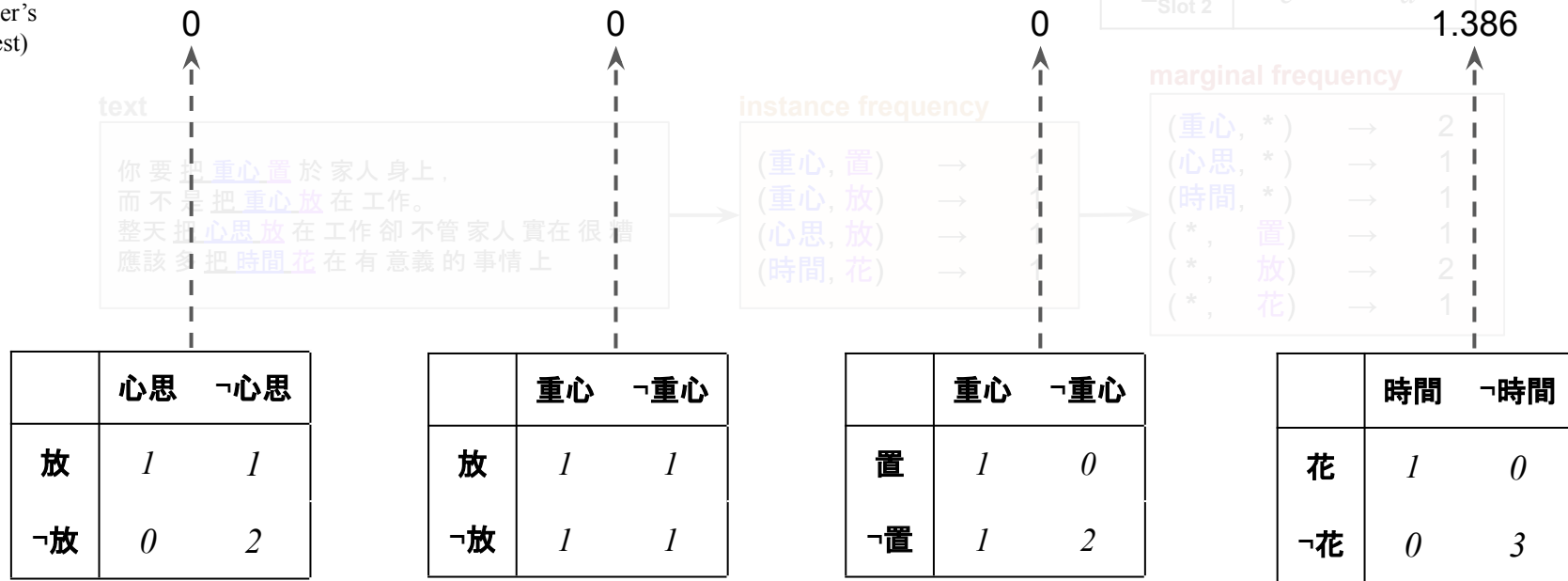
# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個 lexical slots 內的詞彙的共現傾向

e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

Attraction  
(by Fisher's  
exact test)



Attraction for each (N, V) pair is calculated by  $-\log(p)$ , where  $p$  is the  $p$ -value of a Fisher's exact test performed on the contingency table

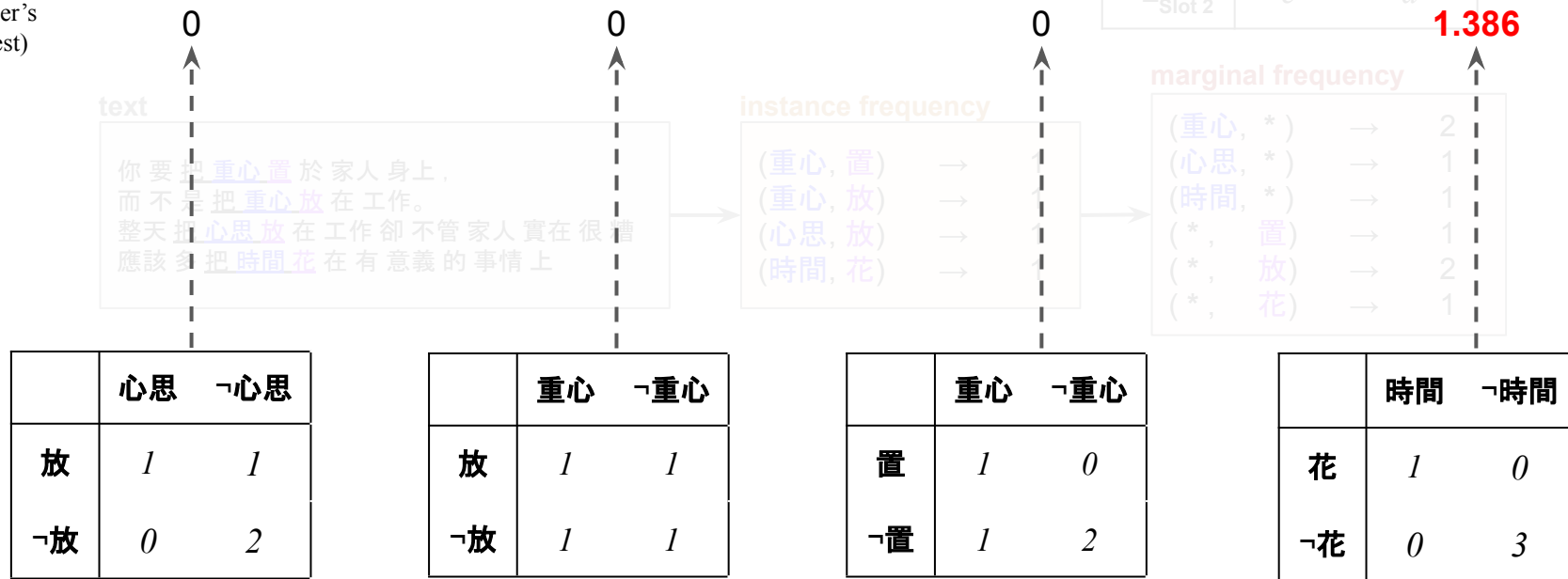
# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個 lexical slots 內的詞彙的共現傾向

e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

Attraction  
(by Fisher's  
exact test)



Attraction for each (N, V) pair is calculated by  $-\log(p)$ , where  $p$  is the  $p$ -value of a Fisher's exact test performed on the contingency table

# Co-varying Collexeme Analysis: toy example

- Co-varying collexeme analysis (Stefanowitsch & Gries, 2005)

- 衡量同一句式下的兩個 lexical slots 內的詞彙的共現傾向

e.g., 「把」字句中的 N 與 V, 如: 把 時間(slot1) 花(slot2) 在...

Attraction  
(by Fisher's  
exact test)

0

0

0

	L <sub>Slot 1</sub>	¬L <sub>Slot 1</sub>
L <sub>Slot 2</sub>	a	b
¬L <sub>Slot 2</sub>	c	d

1.386

text

你要把 重心 置 於家人身上，  
而不是把 重心 放 在工作。  
整天把 心思 放 在工作卻不管家人實在很糟  
應該多把 時間 花 在有意義的事情上

instance frequency

(重心, 置)	→	1
(重心, 放)	→	1
(心思, 放)	→	1
(時間, 花)	→	1

marginal frequency

(重心, *)	→	2
(心思, *)	→	1
(時間, *)	→	1
(*, 置)	→	1
(*, 放)	→	2
(*, 花)	→	1

	心思	¬心思
放	1	1
¬放	0	2

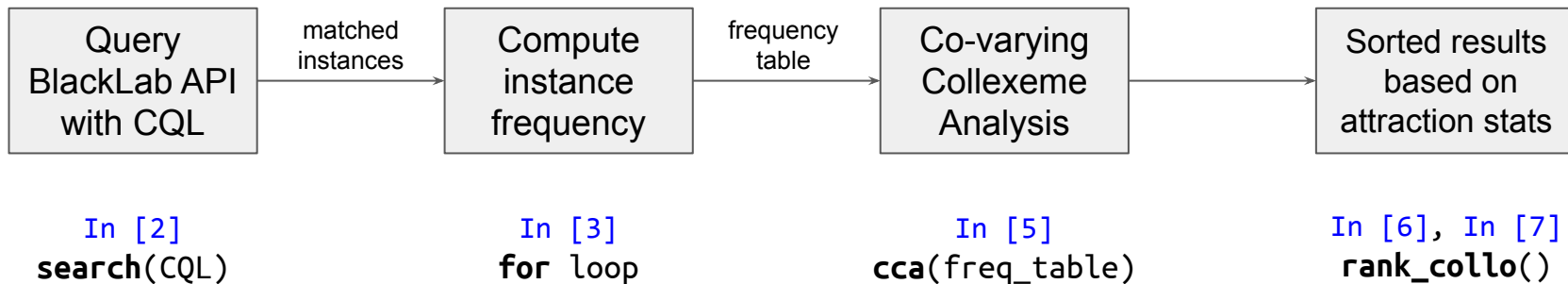
	重心	¬重心
放	1	1
¬放	1	1

	重心	¬重心
置	1	0
¬置	1	2

	時間	¬時間
花	1	0
¬花	0	3

Attraction for each (N, V) pair is calculated by  $-\log(p)$ , where  $p$  is the  $p$ -value of a Fisher's exact test performed on the contingency table

# Co-varying Collexeme Analysis in Python



1. Compute contingency table

2. Compute attraction stats  
(association measures)

<sup>1</sup> GitHub repo: <https://github.com/lorentu/hocor2020-GramColl>

<sup>2</sup> 對照 `collostructional_analysis.ipynb`: “1. Covarying Collexeme Analysis (CCA)” ([view on web](#))

<sup>3</sup> `search()`, `cca()`, `rank_collo()` 說明文件見 <https://lorentu.github.io/hocor2020-GramColl>, 程式碼見 `APIsearch.py` 與 `collo_measures.py`

# References

- Desagulier, G. (2017). *Corpus Linguistics and Statistics with R*. Springer. Retrieved from <https://doi.org/10.1007/978-3-319-64572-8>
- Gries, S. T., & Stefanowitsch, A. (2004). Extending collocation analysis: A corpus-based perspective on alternations'. *International Journal of Corpus Linguistics*, 9(1), 97–129.
- Huang, C.-R., Kilgarriff, A., Wu, Y., Chiu, C.-M., Smith, S., Rychlý, P., ... Chen, K.-J. (2005). Chinese Sketch Engine and the extraction of grammatical collocations. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.
- Stefanowitsch, A., & Gries, S. T. (2003). Collocations: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243.
- Stefanowitsch, A., & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, 1(1), 1–43.