構式之無所不在：

認知與計算觀點 (Version: 0.2)

謝舒凱 (Graduate Institute of Linguistics, NTU)
2021-03-05

Language and Ontology Processing | 語言即計算

語言學有理論嗎 🙈

- THE LEXICON is THE MATRIX ㊙️：需要想像一個終極的研究對象
  - 構詞、詞意區分、概念知識本體、變異與變遷、情意處理與社會網路
- 語言研究的時代精神 Zeitgeist：
  - 社群媒體促生不同語言版本的民主化

    歷史文本到言談多模溝通
  - 語言、資訊與實在：規約設計 (rule/constraint-based) 到使用觀察 (usage-based)

    原子到量子
  - 人際到人機互動

    符碼規則到向量空間

Plan

- Introduction to `construction.grammar`
- Corpus-based (association) measures
- Computational construction grammar
- Conclusion and on-going works

Introduction

Language in use | 生活中的 (新) 語言使用

(Goldberg, 2019)

TABLE 1.1. Novel linguistic exemplars that demonstrate the productivity of various constructions

"Hey man, *bust me some fries*."	Double-object construction
"Can we *vulture your table*?"	Transitive causative construction
"Vernon *tweeted to say she doesn't like us*."	*To* infinitive construction
"What a *bodacious thing* to say."	Attributive modification construction

TABLE 1.2. Novel formulations that are judged odd by native speakers

?She explained him the story. (cf. She told/guaranteed him the story.)	Double-object construction
?He vanished the rabbit. (cf. He hid/banished the rabbit.)	Transitive causative construction
?She considered to say something. (cf. She hoped/planned to say something.)	*To* infinitive construction
?The asleep boy (cf. The astute/sleepy boy)	Attributive modification construction

- The **explain me this** puzzle: USAGE-BASED CONSTRUCTIONIST APPROACH to language.

先從一個「圖文不符」的例子講起

怎麼說一個人帥?

X 成這樣是要怎麼辦啦、 要不要這麼 X 啦、 有 XX 的感覺、 X 的不要不要的

語言表達單位的「連續」

自由語到熟語

- 單字詞 ("`愛`")、雙字詞 ("`愛情`")、三字詞 ("`大不了`")、四字格 ("`沒大沒小`")、四字成語 ("`一葉知秋`")、格言 ("`滿招損謙受益`")、歇後語、諺語 。。

Formulaic Language (語式) | Review

- *Formulae* : 以常用語反覆為特徵，心理上具有預製 (prefabbed) 的語言現象。
- 60 多種相關術語： `定式語` (formulaic sequence/utterance/speech)、`語塊` (chunk)、`詞串` (lexical bundles)、`多詞單位` (multiword expressions)、`搭配詞` (collocation)、`成語` (idioms)、`固定語` (fixed expressions) .....

> a sequence, continuous or discontinuous, of words or other meaningelements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (wray, 2000)

語式的特徵 | Basic characterization of formulaic

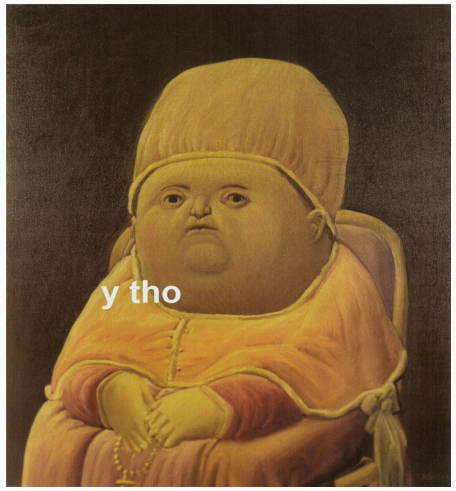
- 語音連貫性
- 語意組合性（`透明到晦澀`）
- 整體提取性
- 句法常常不合常規
- 使用頻率穩定

Construction grammar (C*G) | 構式文法

- 語式的觀察沒有要探究認知意義，構式文法的理論企圖較大。

> Construction grammar (often abbreviated CxG) is a sociobiological family of theories within the field of cognitive and evolutionary linguistics. These posit that human language consists of constructions, or learned pairings of linguistic forms with meanings. (wiki)
>
> C is a CONSTRUCTION iff def C is a form–meaning pair $< F_i, S_i >$ such that some aspect of F_i or some aspect of S_i is not strictly predictable from C's component parts or from other previously established constructions (Goldberg, 1995).
>
> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency (Goldberg, 2006).

Raison d'être



The common idea is that a speaker's knowledge of his/her language consists of a very large inventory of constructions, where a construction is understood to be of any size and abstractness, from a single word to some grammatical aspect of a sentence , such as its Subject-Predicate structure.

CONTRUCTICON

Variations and Basic Tenets

不同版本之共享前提

- different versions (Goldberg; Langacker; Croft....) ; a family of theories.
- Most constructional approaches agree on three basic tenets:

✅ constructions are symbolic units
✅ there is a continuum between lexicon and grammar
✅ constructions are organized in networks.

Constructional Meaning | 意義是建構的

哲學與認知前提

- bottom-up, usage-based 浮現語法觀

> emphasizes that exemplars— structured representations— cluster within a hyper-dimensional conceptual space giving rise to emergent constructions, which are then extendable as needed for the purpose of communication.

- monotonic, non-derivational (thus no transformation rule, etc).

  單層級，不推導。句法現象不是由生成規則運算而生的副產品（如被動不是由主動推導出來）

(I) Form-Meaning Pairings

「形意對」的符碼單位

- The formal aspect of a construction is typically described as a syntactic template, but the form covers more than just syntax, as it also involves phonological aspects, such as prosody and intonation. The content covers semantic as well as pragmatic meaning.

- 大部分的構式具有（不同程度的）「不可預測性」的特質。（如：【X 什麼 X】具有否定、斥責）。成分與構式之間的規約與語意激發關係值得探究。

(II) Syntax-lexicon continuum

「非模組性」打破語言學知識分界

- collapse the classical distinctions between semantics, syntax and pragmatics (or, treated holistically.)

  每個語法 chunk 都是形式、意義/用法的整體 (Gestalt)
- construction grammarians argue that all pairings of form and meaning are constructions including phrase structures, idioms, words and even morphemes.

  涵蓋從語素到句子各級句法單位，甚至詞類的構式

> They laughed him out of the room. (town/court)

It means that they laughed at you until you were so embarrassed that you left the room and kept laughing as you left. > the normally intransitive verb receives a transitive reading and the situation can be interpreted on the basis of the 'X cause Y to move' construction rather than the syntactic deviance alone.

(III) Cognitive Organization

- Grammar as a taxonomic inventory of constructions, which are based on the same principles as those of the conceptual categories known from cognitive linguistics, such as
  - inheritance
  - prototypicality
  - extensions
  - multiple parenting

訊息的儲存、組織與表徵

different models are proposed in relation to how information is stored in the taxonomies.

- full-entry/usage-based/default inheritance/complete inheritance model ……
- since the late nineties there has been a shift towards a general preference for the usage-based model. The shift towards the usage-based approach in CxG has inspired several the development of corpus-based methodologies of constructional analysis.

> The usage-based model is based on inductive learning, meaning that linguistic knowledge is acquired in a bottom-up manner through use. It allows for redundancy and generalizations, because the language user generalizes over recurring experiences of use.

Reflections

Is C*G a (scientific) theory or a perspective?

- cognitive commitment
  - the structures of language emerge from interrelated patterns of experience, social interaction, and cognitive processes.
- computational commitment

Chinese Examples

Affixoid | 類詞綴

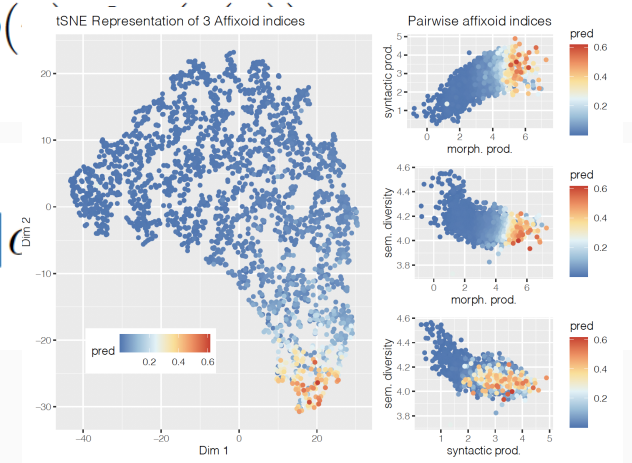
Computational modeling of Affixoid Behaviour in Chinese Morphology

- The morphological status of most Chinese morphemes is indeterminate; rarely derivational and hardly ever inflectional (Hsieh and Huang, forthcoming)
- Modeling affixoid beheviour (Tseng, Hsieh, Chen and Court, 2020)

$$\text{morphological productivity}(\alpha) = a(w_1^\alpha) + a(w_2^\alpha) + \cdots + a(w_N^\alpha) = \sum_{w \in \mathbf{W}^\alpha} \frac{1}{w} \quad (1)$$

$$\text{syntactic productivity}(\alpha) = \sum_{\pi \in \text{POS}(\mathbf{W}^\alpha)} p(\quad (2)$$

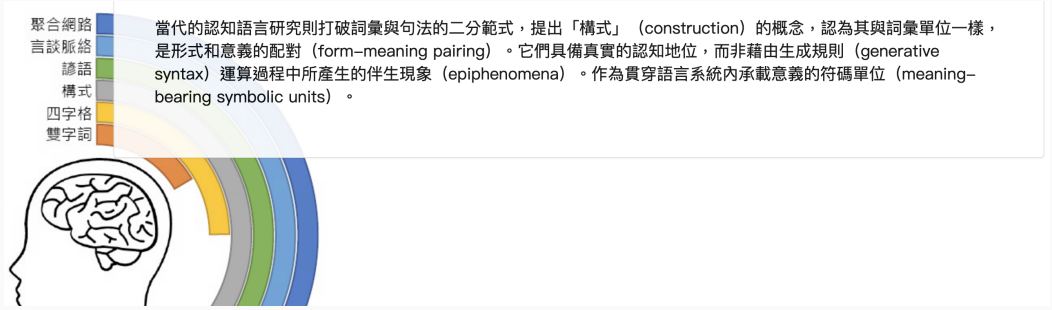
$$p(\mathbf{w}|\alpha, \beta) = \int_\theta \left(\prod_{n=1}^{N} p(w_n|\beta, \theta) \right) p(\theta|\alpha \quad (3)$$



QIEs and Idioms | 四字格與成語

華語四字格詞彙網路的認知神經研究 2017-8

- Prefabs 體現了構式的特點
  - 例：`X頭X尾`：虎頭蛇尾、沒頭沒尾、有頭沒尾、徹頭徹尾、從頭到尾、街頭巷尾
- 結構對詞彙給出了限制



(Hsieh et al, 2017), *Entrenchment and Creativity in Chinese Quadrasyllabic Idiomatic Expressions*

QIEs and Neural Evidence

- We hypothesized that when primed with idioms (and other prefabs QIEs) whose construction are ‘entrenched’, language speakers should be faster in comprehension task, and when the QIEs are not recognizable as fixed units, semantically related targets within the constraints set by construction do play a determining role in the acceptance of QIEs variants.

Construction in opinionated textsAspect-based

> Opinionated text can be defined as the text acquired from blogs, social networking sites or any other online portal in which the users have expressed their disposition and point of view towards any particular product or service.

- 語言是情意表達的戰場,實例

中華電信輿情分析 2017-8; 2020-21

Internet Meme and Multimodal C*G

- **Internet Memes** provide a growing volume of multimodal data. (cf. `Multimodal Meme Dataset`(Suryawanshi et al. 2020))
- Typical image macro includes 3 element:
  - a background **image**
  - top **text** that often *formulaic* and easily recognizable
  - bottom text that often delivers the *punch line* of the theme

Internet Meme and Multimodal C*G

- Our preliminary computational modeling results (Lin and Hsieh, 2020) shows that
  - IMs are multimodal constructions in nature.
  - Different modalities (verbal-visual incongruity) demonstrates differences in *saliency* (or *Figure/Ground* contrast).

- Multimodal Construction Grammar
  (Zima and Bergs 2017; Zenner and Geeraerts 2018)

Modalities	ACC
Captions (OCR)	0.72
Template Names (TM)	0.68
ResNet50 (RN)	0.66
OCR+TM	**0.76**
OCR+TM+RN	0.75

Table 2: Classification results of Binary MLP Classifier

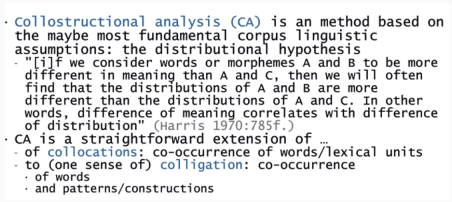
Corpus-based (association) measures

搭配 collogation、搭構 collostruction 與構式

兩個德國 Stefan (Evert and Gries)

- Collocations as a linguistic epiphenomenon: collocation statistics 綜整
- collostructional analysis 介紹

Collostructional analysis (Gries, 2020)



在構式語法 (construction grammar) 架構下，Stefanowitsch and Gries (2003) 發展一套分析法稱 搭構分析 (collostructional analysis)

- Collexeme analysis measures the mutual attraction of lexemes and constructions.
- Distinctive collexeme analysis contrasts alternating constructions in their respective collocational preferences (Gries and Stefanowitsch 2004b).
- To measure the association between the two slots of the same construction, a third method known as covarying-collexeme analysis is used (Gries and Stefanowitsch 2004a; Stefanowitsch and Gries 2005).

Collostruction analysis in practice

Collostructional analysis

- CA is a 'family' of 3 methods
  - collexeme analysis
    - co-occurrence of each of *n* words
    - in/with 1 construction
  - distinctive collexeme analysis
    - co-occurrence of each of *n* words
    - in/with 2 (or more) constructions
  - co-varying collexeme analysis
    - co-occurrence of words in 2 slots of 1 construction
- for each 2x2 table, one computes an assoc. measure to see
  - which words like cx 1
  - which words prefer which cx
  - which words go together in cx
  - based on the (ranks of) sorted association measures (AMs)
- AMs most widely used:
  - $p_{\text{Fisher-Yates exact}}$ & G^2/LLR

	cx 1: y	cx 1: n	Σ
word 1: y	80	200	280
word 1: n	1000	…	…
Σ	1080	…	Σ

&

	cx 1: y	cx 1: n	Σ
word 2: y	60	310	370
word 2: n	1020	…	…
Σ	1080	…	Σ

	cx 1	cx 2	Σ
word 1: y	150	80	230
word 1: n	930	720	1650
Σ	1080	800	1880

&

	cx 1	cx 2	Σ
word 2: y	60	310	370
word 2: n	1020	490	1510
Σ	1080	800	1880

	word 2: y	word2: n	Σ
word 1: y	40	240	280
word 1: n	330	470	800
Σ	370	710	1080

&

	word 3: y	word3: n	Σ
word 1: y	20	260	280
word 1: n	180	620	800
Σ	200	880	1080

公開連結

(slide credit: 廖永賦)

- 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

	L_j	$\neg L_j$
C	*a*	*b*
¬C	*c*	*d*

- 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

	L_j	$\neg L_j$
C_1	*a*	*b*
C_2	*c*	*d*

- 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...

	$L_{Slot\ 1}$	$\neg L_{Slot\ 1}$
$L_{Slot\ 2}$	*a*	*b*
$\neg L_{Slot\ 2}$	*c*	*d*

PTT 構式分析舉例

- 「實在是」、「不能不說是」、「最好是」、..... 「可以再 XX (一點)」「還在那邊 XX」 、「XX 不意外」、「是在 XX」
- 假設我們想研究`「可以再 XX (一點)」`的(負面)語意韻/情感含義，大家會怎麼做?
- R 實作參考; Python 實作參考

Interim summary

- 構式語法與當代認知科學發現的關聯應該是正相關
- 語料庫語言學做出了共現現象的計量觀察與顯著檢定嘗試（至少在經驗基礎上提供了 Exploratory Data Analysis 的計算環境）

但，如果構式不僅是作為一種 add-on 的語言(處理)資源，而是提供理論的基礎時，計算構式 (CoCoGram) 應該如何發生?

計算構式 (CoCoGram) 先導問題

- `如何表徵符碼與意義?` "The form and content are symbolically linked" (form-meaning pairs) but how can both to be co-represented?
- `有無自然邊界？如何調節更新流動與變異？` are texts possible to be naturally chunked? reconcile varieties? roles in the understanding of the text/language?
- `隱性知識本體如何紬繹與回饋？` how a 'provisional unit boundary' can be detected, the 'linear unit of meaning' (linear unit grammar) be acquired and latent hierarchy can be discerned in the linear string of word forms?
- `如何解決同一性`

Computational (Lexical) Semantics

- 語言理論無法忽略語意
- 語言理論要能解釋 (Pustejovsky, 1996):
  - polymorphism
  - semanticality
  - creative use
  - co-composition

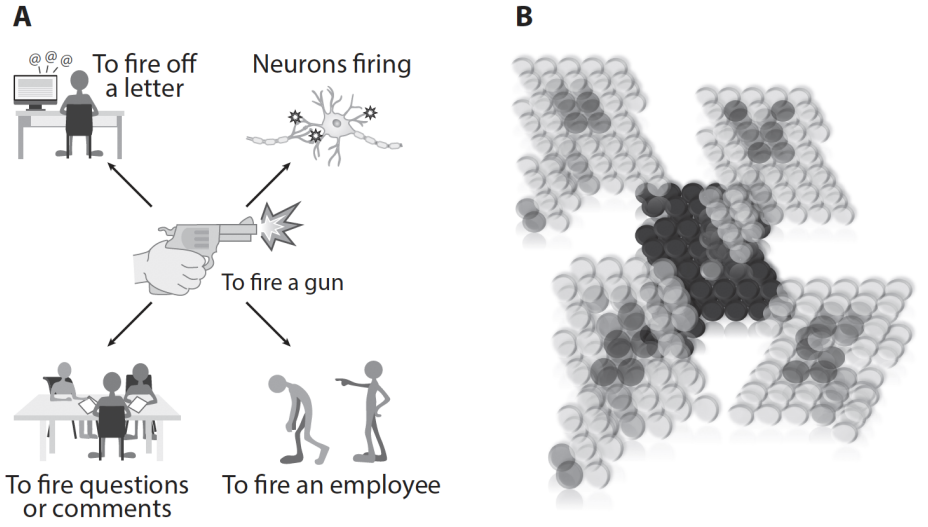


FIGURE 2.3. *A*, symbolic, and *B*, abstract, structured, distributed representation of related senses of the verb *to fire*.

(Goldberg, 2019)

Computational construction grammar (and NLU)

COCOGram aims to operationalize insights and analyses from construction grammar as computational processing models.

It does not only allow for automatic validation of the preciseness and consistency of construction grammar models, and to run these models on text corpora, but also to make use of construction grammar insights to enhance the performance of language technology applications.

- 機器學習與驗證是關鍵

想法

- **Fluid Construction Grammar (FCG)**
  (Remi van Trijp, 2020; https://www.youtube.com/watch?v=YTKHllV4MCU)
  - traditional linguistic processing : vertically eat layer-by-layer >> pipelined NLP
  - constructionist > horizontally eat all
- **Construction Grammar Induction**
  - Computational learning of construction grammar (Dunn, 2017)

Conclusion and Future/on-going works

Conclusion

- 語料庫量度成就了EDA，在漢語與台灣南島語言上可以多做比較
- 認知計算構式方才開始，符碼序列無邊界線索的語言文字提供最難與最關鍵的養分
  - (Symbolic) form and (Sub-symbolic) meaning representation
  - CG induction algorithm

On-going works

- 構式與變異：理論跟著語料（語言使用）走

> 他哭奶奶/小孩就活個媽媽/飛上海/跑生意/跑百米/走八卦掌/喊一聲嗓子/闖紅燈 (郭繼懋 1999)

- Sense-aware Construction/Collostructional Analysis
  Chinese Sense Tagger

Reference

- Bergen, Benjamin & Nancy Chang (2005). "Embodied Construction Grammar in Simulation-Based Language Understanding". In J.-O. Östman & M. Fried, ed. Construction Grammar(s): Cognitive and Cross-Language Dimensions. Amsterdam: John Benjamins. .
- Croft, William A (2001). Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford: Oxford University Press.
- Croft, William A. & D. Alan Cruse (200). Cognitive Linguistics. Cambridge: Cambridge University Press.
- Fillmore, Charles, Paul Kay and Catherine O'Connor (1988). "Regularity and idiomaticity in grammatical constructions: the case of let alone". Language, 64. 501-38.
- Goldberg, Adele (1995). Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.
- Goldberg, Adele (2006). Constructions at Work: the nature of generalization in language. Oxford: Oxford University Press.
- Goldberg, A. (2019). *Explain me this: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton University Press.

Reference II

- Hsieh, S.K. et al. (2017). *cognitive neurological base of idiomatic network in chinese*. ICCS.
- Lakoff, George (1987). Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago: CSLI.
- Langacker, Ronald (1987, 1991). Foundations of Cognitive Grammar. Vols I-II. Stanford: Stanford University Press.
- Michaelis, Laura A. & Josef Ruppenhofer (2001). Beyond Alternations: A Construction-based Account of the Applicative Construction in German. Stanford: CSLI.