# Table of Contents

# List of Figures

# List of Tables

# 歷時語料庫

早在**sinclair1982reflections**，語言學家**sinclair1982reflections**如此描繪了對於未來語料庫模樣的想像，文字的保存是大量的，其中是緩慢卻不斷變動的語料（“vast, slowly changing stores of text”），亦是對語言演化很詳細的紀錄（“detailed evidence of language evolution”）。

語言，將所思所想傳遞、紀錄，並在說話者使用語言時，不斷被重塑與流傳 **(blank1999new)**。從共時（synchronic）的角度來看，語意存在各種變異（variation），而在歷時（diachronic）的脈絡下，經過時間累積而則彰顯了各種的變遷。近年來的歷史詞彙語意研究，從詞意的改變、新舊字詞的興衰，探索其背後的運作機制與認知層面，已開始摸索出語意變遷（semantic change）的規律性（regularities）**(blank1999new)**。

從語料量化與計算的觀點切入詞彙語意變遷的語言現象，近年來文字在網路上大量流傳，加上社會快速變遷，語意表達亦不斷變化。與此同時，歷史文本的電子化數量的增長，使我們得以從中分析、挖掘詞彙所蘊含的詞意，開展了更多與歷時語意相關的研究可能。

語料庫作為語言使用的經驗素材，提供了我們從中觀察、歸納出可質化、量化的語言分析；而歷時語料庫更因應科技進步，結合了計算語言學界近年來的語言向量表徵、神經語言統計模型等新方式探求語意在時間洪流下的變動與趨勢。

(1) Corpus of Historical American English (COHA, 1810-2010)[1]

(2) A Representative Corpus of Historical English Registers (ARCHER, 1600-1999)[2]

(3) Royal Society Corpus (RSC, 1665-1869)[3]

(4) Corpus of Late Modern English Texts (CLMET, 1710-1920)[4]

(5) Hansard Corpus (1803-2005)[5]

(6) Sheffield Corpus of Chinese[6]

(7) Academia Sinica Tagged Corpus of Old Chinese (中央研究院上古漢語語料庫, from pre-Qing to pre-Han)[7], Academia Sinica Tagged Corpus of Middle Chinese (中央研究院中古漢語語料庫, from late-Han to the Six Dynasties)[8], and Academia Sinica Tagged Corpus of Early Mandarin Chinese (中央研究院近代漢語語料庫, from Tang to Qing)[9]. The division into 3 corpora is based on the development of Chinese syntax to offer a synchronic sketch of Chinese and a basis for diachronic comparisons. In the 3 Academia Sinica tagged corpora, raw texts are available, with part of the texts imported from Scripta Sinica (漢籍全文資料庫計畫). It is also worth noting that the Google Books project for Chinese is not available until the year of 1950, and the latest date is 2008. It is believed that corpora creation is the foundation for a more thorough and accurate depiction for data collection during the establishment of lexical databases.

然而在歷時語料中，有些詞彙並無明顯的詞頻變化，其多義行為亦造

[1]https://www.english-corpora.org/coha/
[2]https://www.projects.alc.manchester.ac.uk/archer/
[3]https://fedora.clarin-d.uni-saarland.de/rsc/
[4]https://perswww.kuleuven.be/~u0044428/
[5]https://www.english-corpora.org/hansard/
[6]https://www.dhi.ac.uk/scc/
[7]http://lingcorpus.iis.sinica.edu.tw/ancient/
[8]http://lingcorpus.iis.sinica.edu.tw/middle/
[9]http://lingcorpus.iis.sinica.edu.tw/early/

成研究者面對巨量資料時的困擾。本論文的目的，在於結合語料統計模型與計算語意學的表徵模型，探究漢語的語意變遷。從數位化的原始語料中，以共現（co-occurrence）分佈的趨勢發覺意義分布的異同，並從語境詞向量（contextualized word embeddings）將多義性（polysemy）的變動做形式表達。期待以量化的方式量測語意變遷的程度，並以質化分析輔證已知的例子，並發掘更多可能的例子與規律。我們以歷時語料庫（中國哲學書電子計畫 (**sturgeon2019ctext**)）與現代漢語語料庫（中研院漢語平衡語料庫 (**chen1996sinica**)）為語料來源，建立歷時詞向量並搭配詞彙資料庫，並參考 **hamilton2016cultural** 的全域鄰近詞法，以搭配詞的相似度數值組成二階向量（second-order embedding），提高語意表徵的精確度來比較各時代向量的方法，求其相關係數和語意變遷程度之間的關聯。並從詞彙的意義分布與互動，描繪出不同詞意的消長與變動。此外，本研究也同時採用以變異程度為基礎的近鄰群聚分析法（Variability-based Neighbor Clustering, VNC）(**gries2012variability**)，此階層式的分群可勾勒出綜合性評估各觀察變項的影響下，漢語詞彙發展的時代區分。

計算語意學與歷史語意學的整合研究可以使我們在經驗基礎上回溯驗證個別詞彙的意義變化，更進一步梳理整體的原理原則。詞彙反映人們對於新事物賦予新名的動機、社會概念的更迭也同時牽動詞彙之間的關聯，其應用範圍更可擴及到詞彙與文化變遷的探索。

## 0.1 第一章語言的本質與語言學研究

黃宣範

國立臺灣大學語言學研究所

- – 前言

  – 語言的本質 [FF1A?](**Hockett1960**) 的看法

- 語言是複雜而具調適力的系統
- 語言既非自然物也非人造物
- 投射與語言的時間性
- 語言的複雜性
- 語言的規律性
- 語言的社會性
- 語言與互動引擎 [FF1A?] 大腦是個喜愛預測的器官
- 語言學簡史
- 語言學的次領域
- 語言學在台灣
- 摘要與結論
- 參考文獻

Science enriches us by bringing us beauty in multiple forms

- Elizabeth Blackburn, 2009 Nobel Laureate quoted in New York Times 12/07/2019

Man acts as though they were the shaper and master of language while in fact language remains the master of man -Martin Heidegger

### 0.1.1 1.1 前言

現代智人大約在 10 萬 ~15 萬年前創發了語言 [FF0C?] 但人類把語言透過遷徙傳播到世界各個角落則是很晚近的事。根據 ETHNOLOGUE (ethnologue.com) 的調查 [FF0C?] 迄 2020 年初為止 [FF0C?] 世界上目前使用的語言達 7117 個 [FF0C?] 分別屬於至少 242 個不同的語系 [FF0C?] 這還不包括早已消失的語言。專家估計在

17 世紀帝國主義興起之前 [FF0C?] 世界上的語言可能多達 15000 個。世界上的語言不但多 [FF0C?] 而且由於語系多 [FF0C?] 彼此岐異非常大 [FF0C?] 結構上的差異極為懸殊。有些語系包括的語言特別多 [FF0C?] 例如南島語系涵蓋約 1200 個語言 [FF1B?] 有些語系則只有一個語言 [FF0C?] 就是所謂孤立語 [FF0C?] 自成一格 [FF0C?] 不屬於任何語系。例如西班牙與法國邊界的巴斯克語 (Basque)[FF0C?] 或北海道的愛奴語 (Ainu)。

這 7117 個語言在地球的分布差異頗大。有兩個國家語言分布的密度特別高: 巴布亞新幾內亞 [FF0C?] 面積約台灣 13 倍 [FF0C?] 但有 851 個語言; 也同樣是南太平洋的萬那杜 [FF0C?] 面積是台灣的三分之一 [FF0C?] 但有 115 個語言。根據 (**Nettle1999**) 的統計 [FF0C?] 幾個大洲之中 [FF0C?] 非洲使用的語言最多 [FF0C?] 因為非洲是人類的語言的原鄉 [FF1B?] 其次是南亞與東南亞 [FF0C?] 再其次是新幾內亞與南美洲。其實整個美洲合起來也超過 1200 個語言 [FF0C?] 僅次於非洲以及南亞東南亞 [FF0C?] 如下表所列 [FF1A?]

表 1-1 語言在地球的分布 (根據 (**Nettle1999**))

| | |
|---|---|
| 非洲 | 2614 |
| 北歐亞大陸 | 732 |
| 南亞東南亞 | 1998 |
| 大洋洲 | 306 |
| 新幾內亞 | 1109 |
| 澳洲 | 234 |
| 北美洲 | 243 |
| 中美洲 | 381 |
| 南美洲 | 595 |

細心的讀者可能特地把上面的語言加總起來 [FF0C?] 發現超過 7117 個。為甚麼 [FF1F?] 至少有兩個因素。其一 [FF0C?] 上面的表是根據九個大洲/次大洲來劃

分 [FF0C?] 所以英語 [FF0C?] 西班牙語 [FF0C?] 法語等強勢語言就會重複計算 [FF0C?] 因為這些語言在好幾個國家都使用。其二 [FF0C?](**Nettle1999**) 的統計也是根據 ETHNOLOGUE 的原始資料 [FF0C?] 但 ETHNOLOGUE 隨時都在修正更新 [FF0C?] 舊的版本難免錯誤較多 [FF0C?] 例如有些語言可能已經消失 [FF0C?] 但仍列入計算 [FF09?]。

一般人馬上會問 [FF1A?] 怎麼才算是一個語言 [FF0C?] 而不是方言 [FF1F?] 這有常人的想法 [FF0C?] 也有專家的看法。最簡單的回答是 [FF1A?] 如果對方的話你基本上聽不懂 [FF0C?] 那麼這個語言就是另外一個語言 [FF1B?] 反之 [FF0C?] 你基本上聽得懂對方八成以上 [FF0C?] 那麼那個語言大概跟你講的話是同一個語言 [FF0C?] 可能只是個方音詞彙有些微的差別而已。華語跟客家話 [FF0C?] 雖然都是漢語 [FF0C?] 其實是兩個不同的語言 [FF0C?] 因為彼此無法溝通 [FF0C?] 所以不僅僅是不同的方言而已。中國一向有八大方言之說 [FF0C?] 但這八大方言彼此無法溝通 [FF0C?] 卻由於源自同一個古語 [FF0C?] 又都以漢字書寫 [FF0C?] 因此官方定調為方言的。但多數學者的理解將之歸為八個語言 [FF0C?] 或視為連續體的方言鍊。方言鍊可能存在人口多地域又廣的語言 [FF1B?] 中國大陸的普通話幅員廣及四川到天津 [FF0C?] 方言的差異頗大 [FF0C?] 兩個極端很可能也不易溝通。台灣南島語的賽德克語跟太魯閣語根據我們做過的調查 [FF0C?] 彼此溝通無礙 [FF0C?] 自當視為同一個語言。在原來日本學者分類系統下的 Sejiq[FF08?] 現在一般拼音為 Seediq[FF09?] 語支的原住民族中 [FF0C?] 後來「太魯閣地區」的族人發展出自己的文化認同 [FF0C?] 傾向以「太魯閣族」作為族群名稱。而南投地區的「德路固」[FF08?]Truku[FF09?]、「德克達雅」[FF08?]Tgdaya[FF09?] 及「道澤」[FF08?]Toda[FF09?] 的族人傾向以「賽德克族」作為族群名稱 [FF0C?] 因而展開了族群內部的衝突與對話。最後在 2004 年太魯閣族獨立成為官方認定的一個族群 [FF0C?] 隨後賽德克族也在 2007 年獨立為另一個族群 [FF0C?] 兩個族語也就被

官方認定為兩個不同的語言。這個故事顯示社群/族群的意識多少決定語言的歸類。

在外國類似的例子也不少。興地語 [FF08?]Hindi[FF09?] 跟烏爾都語 [FF08?]Urdu[FF09?] 都是印度主要語言 [FF0C?] 前者是印度官方語言 [FF0C?] 後者剛好也是巴基斯坦的官方語言 [FF0C?] 但這兩個語言其實溝通沒有問題 [FF0C?] 唯一的不同在於書寫系統 [FF1A?] 印地語用的是天城文字 [FF08?]devanagari[FF09?][FF0C?] 烏爾都語用的是阿拉伯文字。書寫系統的不同導致學者將之歸類為兩個不同的語言。同樣 [FF0C?] 丹麥話跟挪威話其實是同一個語言 [FF0C?] 但因分屬兩個不同的國家 [FF0C?] 因此被視為為兩個語言。由此可以想見上文提到的 7117 個語言不會單純的只是學理上的分類 [FF0C?] 有時候也涉及政治文化意識的考量。語言學文獻上偶而會看到這樣的比喻 [FF1A?]**A language is a dialect with an army and navy**"[FF08?] 語言是擁有軍隊的方言 [FF09?][FF0C?] 意在顯示語言/方言的區分其實有時候相當武斷。有趣的是每個人生下來 [FF0C?] 面對這些岐異的語言 [FF0C?] 結構如此懸殊 [FF0C?] 但只要給他機會 [FF0C?] 都可以輕易地習得任何一個或多個語言 [FF0C?] 變成雙聲帶或多聲帶。對研究神經心理語言學的學者而言 [FF0C?] 一個重要的議題是 [FF1A?] 到底是甚麼樣的神經心理機制使人類能有如此奧妙的語言習得能力 [FF1F?] 這個議題一向有兩派不同的想法。有一派學者走的是杭士基的路線 [FF0C?] 認為語言的習得這個「柏拉圖的議題」基本上歸因於語言專屬的天賦的能力 [FF1B?] 另外一派認為語言的習得基本上是個「亞里斯多德」的議題 [FF0C?] 認為語言的結構是兒童從各種語境情境中歸納而確立 [FF0C?] 語言跟其他能力一樣都源自人類廣泛而基本的一套認知策略 [FF08?] 參見第 10 章、12 章 [FF09?]。值得注意的是小孩習得他/她周遭的語言跟人工智慧所提倡的深度學習很不同。人工智慧的深度學習需要給予電腦成千上萬的句子去學習 [FF0C?] 並且要加註豐富的語法標記的句子

[FF1B?] 但小孩學習語言完全不需依靠這樣的語法標記 [FF01?]

研究語言指的是研究這些語言的結構 [FF0C?] 語法的結構 [FF0C?] 詞法的結構 [FF0C?] 語意的結構 [FF0C?] 聲韻的結構 [FF0C?] 跨語言的類型研究 [FF0C?] 嬰兒如何習得語言等等。語言的結構成分有大小不同的單位 [FF0C?] 諸如音位 [FF0C?] 詞素 [FF0C?] 詞組等。音位組成詞素 [FF0C?] 但組成過程往往產生各種有趣的變化 [FF0C?] 變化有其規律 [FF0C?] 也有諸多例外 [FF0C?] 而且不同的語言詞序變化頗大 [FF0C?] 表達細膩而層出不窮的意義 [FF08?] 參見第四章 [FF09?]。語言是人生生活際遇的基本配備 [FF0C?] 因此語言學研究的觸角也廣及人文社會以及認知 [FF0C?] 生命或資訊科學有關的很多領域。本書涵蓋的諸多章節很容易看得出語言學研究領域寬廣的視野。語言學研究意義與推論 (見第四章)[FF0C?] 研究語言如何演變 [FF08?] 見第九章 [FF09?][FF0C?] 研究語言與老化 [FF08?] 見第 14 章 [FF09?][FF0C?] 研究語言跟手勢 [FF08?] 見第 11 章 [FF09?][FF0C?] 研究手語的結構 [FF08?] 見第 13 章 [FF09?][FF0C?] 研究不同的社會階層的語言現象 [FF0C?] 研究多語社會與族群認同或權力結構的關係 [FF08?] 見第八章 [FF09?][FF0C?] 研究語料庫與計算語言學 [FF08?] 見第 15、16 章 [FF09?][FF0C?] 生態語言學 [FF08?] 研究語言與生態環境的關係 [FF09?][FF0C?] 研究法庭語言學 [FF08?] 即語言有關的資訊如何當作法庭上的證據 [FF09?][FF0C?] 研究語言與演化及其腦神經基礎 [FF0C?] 研究臨床語言學等等。典型的語言學工作者或有各種不同的抱負 [FF0C?] 但最大的志業可能是到一個陌生的語言去做田野調查 [FF0C?] 從零開始 [FF0C?] 去發現這個語言的結構 [FF0C?] 並且持續經營多年 [FF0C?] 最後完成一本具有價值的參考語法著作。另外假定你被公司賞識 [FF0C?] 派去非洲或南美洲出差 [FF0C?] 負責拓展業務 [FF0C?] 你被迫從頭開始學當地的語言 [FF0C?] 以求多少有助於公司事業的發展。那個地方講的也是一個完全陌生的語言。你試著從翻譯簡單的單字開始 [FF0C?] 期望能找出一對一的音義的關係 [FF0C?] 憑著一些基本的

語言常識 [FF0C?] 再加上幾分「天賦能力」你也可能很快掌握到這個語言的七八成聽說寫的能力。

無論如何 [FF0C?] 語言學工作者感興趣的是 [FF1A?] 這些眾多的語言是否有共通的特性 [FF1F?] 語言應該跟人類一樣。世界上人口雖多 [FF0C?] 但都是現代智人的後代 [FF0C?] 都有相通的神經生物結構 [FF0C?] 相同的生理學現象。語言樣態雖然豐富多樣 [FF0C?] 但根據過去多年的研究 [FF0C?] 這些不同的語言 [FF0C?] 從某種深層的角度看 [FF0C?] 必有一些基本共通之處。這些共通之處就是本章所謂語言的本質。

### 0.1.2 語言的本質 [FF1A?](Hockett1960) 的看法

Hockett 是 20 世紀中葉美國結構學派的靈魂人物。他的重要貢獻之一是提出對語言本質的看法。(**Hockett1960**) 一文的目的是在區分自然語言跟動物的「溝通」方式的根本差異。Hockett 的看法是自然語言擁有所有他所主張的 14 種設計特徵 [FF08?]design feature[FF09?][FF0C?] 而動物 [FF0C?] 不論是蜜蜂 [FF0C?] 猩猩 [FF0C?] 或鸚鵡等動物則最多只有其中的一小部份的特徵而已 (這 14 個完整的設計特徵請見第 13 章)。這些設計特徵之中本節只討論其中重要的五個 [FF0C?] 顯然為人類的語言獨有 [FF0C?] 絕不見於猩猩的呼叫溝通系統 [FF0C?] 或貓的喵喵叫聲。這五個是 [FF08?]1[FF09?] 遙指能力 [FF08?]displacement[FF09?][FF0C?] 指語言 [FF08?] 人類 [FF09?] 或呼叫 [FF08?] 猩猩 [FF09?] 或喵喵叫聲 [FF08?] 貓 [FF09?] 是否有能力指向任何時空遙遠的事物。答案是 [FF1A?] 語言可以談天南地北的任何事物 [FF0C?] 而猩猩或貓在呼叫或喵喵叫時 [FF0C?] 牠要表達的事物一定就在當前 [FF1B?][FF08?]2[FF09?] 反射能力 [FF08?]reflexiveness[FF09?] 是指語言不但可以談外界的事物 [FF0C?] 也可以談論語言本身。談外在事物的語言稱為物體語言 [FF08?]object language[FF09?][FF1B?] 談論語言本身的語言稱為後設

語言 [FF08?]metalanguage[FF09?]。我們甚至也可以用語言談後設語言 [FF0C?] 例如用語言談論任何理論的議題。你能想像即使再聰明的兩隻猩猩牠們之間有能力「談論」自己剛剛發出的叫聲有否達到警告的效果嗎 [FF1F?] 不可能。(3) 滋生力 (productivity) 是指語言或呼叫系統或貓的是否有能力表達許許多多近乎無窮的信號。答案是 [FF1A?] 語言能表達的意義近乎無窮 [FF0C?] 但猩猩 的呼叫系統就只有有限的幾種。根據 (**NewmanWeitzman2015**) 的研究 [FF0C?] 貓的喵喵叫聲‘詞彙’一共只有 24 種 [FF08?] 雖然貓也能靠著尾巴 [FF0C?] 軀體 [FF0C?] 嘴型等表達各種情緒或慾望。對貓的溝通系統有趣的讀者可參閱 (**Schötz2018**)。1960~80 年代有不少學者投入相當的人力物力研究靈長類動物的溝通能力 [FF0C?] 包括教牠們使用手語等等 [FF0C?] 但結果並沒有否定上面的說法 [FF0C?] 即在自然的生態環境下 [FF0C?] 猩猩的呼叫系統基本上沒有滋生力或其他設計特徵。另外一個特徵是 (4) 傳統傳承 (traditional transmission ): 指兒童是從父母或長者或周遭習得語言 [FF0C?] 每一代都要重新學習自己的語言 [FF0C?] 但代代傳輸下去 [FF1B?] 相形之下 [FF0C?] 猩猩或貓咪的呼叫能力是一種與生俱來的本領 [FF0C?] 不是來自上一代的傳承。最後一個特徵是 [FF08?]5[FF09?] 二元結構 (duality of patterning )[FF0C?] 這是自然語言最重要的特徵指語言含有兩層結構 [FF0C?] 一層是沒有任何意義的個別而且有限的音素 [FF0C?] 例如 b、d、g、p、t、k、i、a、o、u 等 [FF08?] 在手語研究中對應於音素的單位是手形、手掌的朝向、動作、位置等 [FF09?][FF1B?] 另一層是把這些個別的、有限的、無意義的音素組合成為有意義的單詞 [FF0C?] 單詞再組成無窮的句子。任何動物的溝通系統都無法分析出這樣的二元結構 [FF08?] 關於手語更深入的討論 [FF0C?] 請見第 13 章手語語言學 [FF09?]。

除了上述的設計特徵之外 [FF0C?]HOCKETT 沒有提到的一個特徵是自然語言都有否定詞跟疑問詞, 可以以之否定或詢問別人 [FF0C?] 但這絕對不見於動

物之間的溝通。試問: 動物有能力打臉牠的同伴說: 老虎出現了 [FF0C?] 但你沒有發警告叫大家快逃。另外一個語言學界公認的重要的特徵是法國學者索緒爾 [FF08?]Ferdinand de Saussure[FF09?] 最先發現 [FF0C?] 之後也經常被一再強調的 [FF0C?] 那就是語言的任意性。在下面的敘述中我有時改用「數位性」一詞 [FF0C?] 以便與「類比性」形成對比。語言的任意性指的是語言的單字所含的語音跟單字的詞意基本上沒有必然的關係; 而數位性是指語言所有的單字一定是由少數幾個獨立的音位組成 [FF0C?] 音位組成無窮多的詞 [FF0C?] 詞組成無窮多的詞組 [FF0C?] 詞組組成無窮多的句子。語言的任意性俯拾皆是。例如中文的「好棒」的「棒」跟「棒子」的「棒」無關 [FF1B?] 英文的 evening 是‘傍晚’[FF0C?] 也是‘扯平’[FF08?] 動詞 even 加上進行式後綴 -ing[FF09?][FF0C?] 兩個意義無關 [FF0C?] 但拼音一樣 [FF1B?]bat 是蝙蝠 [FF0C?] 也是球棒 [FF0C?] 兩者也毫無關聯。拼音相同 [FF0C?] 純粹是歷史的偶然。布農語 taki 是「糞便」, 也是「居住; 屬於」。兩者表面上毫無關聯 [FF0C?] 但有些學者提出 taki 創生論 [FF0C?] 指出糞便的前身是食物 [FF0C?] 布農人很可能藉著吃吃出了居住的領域家園 [FF0C?] 亦即自己的住居 [FF0C?] 乃至於自然世界的想像 (田哲益 2003)。這可以說是布農族版的「道在屎溺」。

語言的任意性來自語言基本上是約定俗成的系統。17 世紀的莎士比亞也信奉語言的任意性: 茱麗葉在莎士比亞的羅密歐與茱麗葉中這麼說過一句名言 [FF1A?]

- What’s in a name? That which we call a rose by any other name would smell as sweet.

語言的任意性 [FF0C?] 讓我們多少可以自由地用任何音串表達我們要表達的意念 [FF0C?] 而不必考慮這些音串是否直接或間接對應外在的事物。dian ‘點’跟 tian‘舔’只差一個音位的不同 [FF0C?]/d/ 對應於/t/[FF0C?] 有 /d/ 跟沒有/d/ [FF1B?] 有/t/ 跟沒有 /t/的區別 [FF0C?] 而不是 /d/ 多寡的程度 [FF0C?] 也不是 /t/ 多寡的程

度。但 dian ‘點’跟 tian ‘舔’卻是很不同的概念。語言的本質在於基本上它的音串是數位的 [FF0C?] 而不是類比的系統。放眼世界上的語言 [FF0C?] 語言的任意性/數位性其實擴及每一個結構層面 [FF0C?] 從音串到詞串 [FF0C?] 從詞串到句子 [FF0C?] 都有其任意性 [FF0C?] 但在詞彙的層次其數位性/任意性可能最為明顯。法國學者索緒爾 (Saussure) 從語言結構的任意性特質出發 [FF0C?] 了解到語言的單詞具有深刻的符號學意義 [FF0C?] 進而發展他的結構語言學。有些學者更認定最近二三十年成為語言學新的派典的認知語言學其思潮可追自索緒爾的結構語言學的理念。

語言的數位性賦於語言無窮的性格。任何信號系統要能表達無窮的信號必須具有數位性。遺傳學中的密碼子 (codon) 類似語言的單詞 [FF0C?] 單詞由四種鹼基配對組成; 外顯子 (exon) 類似語言的段落 [FF0C?] 基因類似語言的篇章 [FF0C?] 基因體則類似一本書。人體這本書含有 30 億以上的單詞 [FF0C?] 可見基因的無窮性。

嚴格講 [FF0C?] 語言的數位性是漂浮在一個有類比性格的大海上。這是語言學者 Bolinger[FF08?]1968[FF09?] 的觀察。換言之 [FF0C?] 人類的語言兼具數位性與類比性的雙重特質。人累了 [FF0C?] 聲音就比較微弱 [FF1B?] 人個性溫和 [FF0C?] 語氣就比較柔和 [FF1B?] 但人憤怒時 [FF0C?] 聲音就高亢。講話時語氣是否堅定 [FF0C?] 可以看他聲調及句末的語調下降的程度。試比較有人無奈地說 ‘算了’ [FF1B?] 另一方面如果兩個人大吵之後 [FF0C?] 其中一個人說 ‘算了’ [FF0C?] 然後拍桌悻然離去。語言的聲音象徵性 [FF08?]sound symbolism[FF09?] 也說明語言的數位性/任意性的說法事實上有其侷限。語言的聲音象徵性是指語言有時會利用某些語音的不同表達某種象徵意義。文獻上最常見的例子是英語發音時口腔最不張開的前高元音 /i / 往往象徵細小 [FF0C?] 而另一個極端是口腔張的最開的後低元音 /ɑ/ 則象徵大一點的事物。wee /wi/ hours ‘凌晨兩三點’ [FF1B?]teeny /tini/ ‘很小’ [FF1B?]teeny-weeny ‘很小很小’。chip ‘[FF08?] 切成 [FF09?] 細片’ [FF1B?]chop

‘[FF08?] 切成 [FF09?] 大片’。中文有時似乎也用母音 /ɔ / 或 /ɑ/ 表示很大 [FF1A?] 宏 / hɔŋ / 圖大/ /dɑ/展 [FF1B?] 洪 / hɔŋ / 水 [FF1B?] 大 /dɑ/ 紅 / hɔŋ /; 為人海/hɑj/派。

文獻上很有名的日語聲音象徵性現象 (有時統稱擬態語 mimetics)。這一類的擬態語通常當作副詞使用 [FF0C?] 而這些擬態語特別之處在於它不但擬聲 [FF0C?] 也擬態 (如 irakira to hikaru 閃閃發光) [FF0C?] 更擬情 (如 dokidoki suru 心怦怦跳)。英文常見的聲音象徵性例子是英文以 gl- 為首的單字似乎都跟「視覺」有關 [FF1A?]glare、glint、gleam、glitter、glance、glimmer、glimpse。但聲音的象徵性不能過度膨脹解釋 [FF0C?] 否則其數位性/任意性就無法成立了。中文有許多含母音 /ɔ / 的單字都跟大無關 (東、董、動、松、送等)[FF1B?] 同樣 [FF0C?] 英文也有許多以 gl- 為首的字也跟「視覺」扯不上關係 [FF1A?]glucose、glory、globe、glad、gladiator、glutton。對這一類問題感興趣的讀者不難想得到更多的例子。

圖像 **[FF0C?]** 指示 **[FF0C?]** 符號 **[FF1A?]** 人類是唯一能使用符號的靈長類

任何溝通用的信號 (sign) 基本上有三大類 [FF0C?] 即圖像 (icon)[FF0C?] 指示 (index) 與符號 (symbol)。人類的語言跟動物的溝通方式有著巨大無法跨越的鴻溝 [FF0C?] 那就是動物的溝通方式只限於前兩類 [FF0C?] 圖像與指示 [FF1B?] 人類的語言使用的絕大多數是符號。有趣的是 [FF1A?] 圖像或指示兩種信號的使用都是由比較古老的腦幹或邊緣系統所掌控 [FF0C?] 人類的語言則是由大腦的新皮質所掌控。上面提到語言學者 (**Bolinger1968**) 的觀察指出人類的語言是個具數位性格的系統漂浮在類比性格的大海上。動物的溝通 [FF0C?] 不論是呼叫與否 [FF0C?] 都是具有強烈的類比性格 [FF0C?] 亦即具圖像性 (iconic)[FF0C?] 而動物的呼叫系統幾乎完全受制於環境因素 [FF0C?] 因此具有強烈的指示 (indexical) 性格 (指向某一個時間或空間鄰近或有因果關係的外在事物)。人類的語言具有數位性 [FF0C?] 因此超越了類比性格 [FF0C?] 或圖像性 [FF1B?] 人類的語言可以談任何天南地北

任何事物 [FF0C?] 具有遙指能力 [FF1B?] 語言更有反射能力 [FF0C?] 除了可以談論任何外界的事物 [FF08?] 這是語言的第一層功用 [FF09?][FF1B?] 語言也可以輕易地變為後設語言 [FF0C?] 用語言談論語言 [FF0C?] 這是語言的第二層作用。當然我們也可以用語言談後設語言 [FF0C?] 那就是語言的第三層功用。人類的語言之所以有如此後設的能力 [FF0C?] 根本原因乃在於語言是抽象的符號系統。語言的每一個字詞 [FF0C?] 即每個信號 [FF0C?] 都是一種符號。符號的使用不必然跟外在世界是否有任何事物與之直接對應。符號獨立於這樣的對應關聯。跟符號的使用比較有關係的反而是跟其他的符號 [FF0C?] 而非外在的事物。動物溝通系統的指示性是指向外在某些事物 [FF0C?] 如果說語言有什麼指示性可言 [FF0C?] 其指示性是指向其他符號 [FF0C?] 因為符號之所以有意義 [FF0C?] 能外指 [FF0C?] 來自於符號與符號間的關聯 [FF0C?] 就是來自於符號間的語法連結關係。而這些語法關係其實衍生自上面提到的更為基本的類比性格或指示性格 ( Haiman 一書最早注意到語法的類比性格)。這就是所謂語言的符號的紮根問題 (grounding problem)。這個重要的觀點來自 Deacon (1997, 2020)[FF0C?] 一位專攻腦神經生物科學以及認知演化的生物人類學家。Deacon 認為語言的系統是建立在比較基本 [FF0C?] 比較原始的類比性與指示性功能之上 [FF0C?] 人類在大腦演化過程中發現了符號 [FF0C?] 從而一舉跨越了為類比性與指示性所侷限的動物的溝通系統的門檻。

、手勢、肢體動作]1.2.2 語言的多元媒介 [FF1A?] 聲音、手勢、肢體動作

我們上面所謂的語言指的是一般人口說的語言 [FF0C?] 指藉著聲音來傳播收訊的口語。這是大家比較熟悉的語言一詞的用法 [FF0C?] 但手語也是廣義的語言的一種。事實上語言的媒介是多元的 [FF0C?] 包含聲音、包含手勢、及整個肢體動作 [FF08?] 關於手語 [FF0C?] 請見本書第 13 章的討論 [FF1B?] 關於手勢的研究 [FF0C?] 請見第 11 章 [FF09?]。以聲音為媒介的語言有其演化上的優勢 [FF0C?] 而現代智人之所以發展使用聲音為媒介的語言應非意外。有人講話時 [FF0C?] 我們

無須看著他 [FF0C?] 無須靠近他 [FF0C?] 更不必依賴風向 [FF0C?] 就可以聽到他的話 [FF0C?] 聽懂他的話。相形之下 [FF0C?] 許多動物信息的傳播 [FF0C?] 接收者必須看著傳播者 [FF08?] 例如手語 [FF09?][FF0C?] 必須緊貼著傳播者或碰觸到傳播的媒介物 [FF08?] 例如盲人之間使用的點字 [FF09?]。當然 [FF0C?] 最常見的口語是面對面時使用的對話。

一般人會話時不僅僅使用聲音作為媒介 [FF0C?] 也使用許許多多的各式各樣的表情或手勢或動作。有些人肢體語言比較豐富 [FF0C?] 但所有的人或多或少一定都會有肢體語言。有兩種肢體動作值得在這邊提一下 [FF0C?] 以顯示語言跟肢體動作之間關係密切。一種肢體動作完全獨立於語言 [FF0C?] 無須搭配語言就有表情達意的作用 [FF0C?] 很像是單字或成語一樣。揮手表示打招呼或再見 [FF0C?] 聳肩表示不知道 [FF0C?] 不在乎。手掌高舉微彎 [FF0C?] 手指作招呼的動作 [FF0C?] 表示要對方過來。這種手勢在第 11 章稱為表徵手勢。我們冷眼旁觀兩個人在對話時很容易注意到兩個人手勢或臉部表情隨著心情或情緒而有變化。一旦講話激昂時 [FF0C?] 肢體語言就更為豐富 [FF0C?] 尤其更會跟某一個重音節一起搭配出現 [FF0C?] 達到強化的效果。想像有人生氣時說「我絕對不幹」時 [FF0C?] 重音落在「絕對」一詞 [FF0C?] 他的肢體動作也會落在那個詞上面。當你講話遇到困難 [FF0C?] 找不到你要的人名或單字時 [FF0C?] 手會一直在胸前作類似畫圓圈的動作 [FF0C?] 表示一時想不出來 [FF0C?] 似乎有意藉著肢體動作把那個人名或地名「引」出來 [FF01?]

### 0.1.3 1.3 語言是一種複雜且具有調適力的系統

語言是一套系統 [FF0C?] 而且是一種複雜且具有適應力的系統 (complex adaptive system) BecknerEtAl2009。小孩學母語 [FF0C?] 或是我們學第二外語 [FF0C?] 一定是藉著傾聽 [FF0C?] 觀察周遭講這個語言的人講出來的許許多多的話語 [FF0C?]

以及他們講話時的語言行為 [FF0C?] 然後加以歸納 [FF0C?] 推論 [FF0C?] 慢慢的逐漸掌握到語言的底蘊、語言的系統。小孩在學母語的過程中 [FF0C?] 歸納或推論都不是有意識的行為 [FF1B?] 相反的 [FF0C?] 成年人學習第二外語則幾乎無時無刻不在有意識地作記憶推衍或歸納。無論如何 [FF0C?] 大人或小孩 [FF0C?] 一旦學會了一個語言 [FF0C?] 這個系統就深藏在腦海裡 [FF1B?] 語言學者的工作就在利用各種研究方法 [FF0C?] 挖掘這個在腦海裡的系統的樣貌。這個系統現在很多學者認為是一種複雜且具有調適力的系統 [FF1B?] 或生物學者所謂的能「自我組織」(self-organize) 的系統。自我組織指的是從最初的無序狀態中 [FF0C?] 藉著各部分之間的相互作用 [FF0C?] 加上逆向的回饋 [FF0C?] 而產生一個變得有序或很有協調作用的過程。這個過程是自發性產生 [FF0C?] 沒有任何中介主導 [FF0C?] 也沒有任何系統內部或外部的系統在控制。「複雜」是指參與系統的運作的個體非常多 [FF0C?]「調適」是指系統隨時可以變化 [FF0C?] 以適應新的情境。自我組織的系統是個非線性系統 [FF0C?] 因為它整體的特性具有局部成分所沒有的特性。在一個非線性的動態系統中 [FF0C?] 結構通常會越來越趨複雜 [FF1B?] 而且結構的形成不需假借任何外在的中樞指揮居間策畫 [FF0C?] 而是因為語言使用者彼此有相同的目標 [FF0C?] 有同樣的認知策略 [FF0C?] 而且彼此調適 [FF0C?] 最後慢慢形成一個系統。自然界或人文社會的情境有許多複雜且具有適應力的系統。試舉一個比較淺顯的例子 [FF1A?] 假定有許多人在夜市「圍觀」賣膏藥的表演秀。這時候 [FF0C?] 圍觀的人潮之所以用「圍觀」一詞來形容絕非偶然。在這種情況下 [FF0C?] 人潮很自然地會形成半圓形 [FF0C?] 不可能是長方形或梯形。而且人潮之所以形成半圓形當然不會是有人下達命令 [FF0C?] 而是每個人都站在自認較為有利的位置 [FF0C?] 結果人潮就自然而然形成一個半圓形。這種現象就是一種「自我組織」。

法律規範也是一種自我組織 [FF0C?] 尤其是初民社會的律法意識。有人可能以

為法律是一群專家坐在一起 [FF08?] 例如立法院 [FF09?][FF0C?] 相互辯論或依照某些邏輯推論原理 [FF0C?] 推敲出來的一個理性的制度。其實不盡然。法律一開始是基於人類普世的直覺和經驗 [FF0C?] 自然形成的系統。在原住民族中 [FF0C?] 族人共同遵守的祖訓戒律和規範 [FF0C?] 布農族稱為 samu[FF0C?] 泰雅族稱之為 gaga[FF0C?] 賽德克族稱為 gaya。無論是 samu, gaga 或 gaya[FF0C?] 都是廣義的倫理律法最主要的核心觀念。當然 [FF0C?] 現代社會的法律條文確實是經過一再的推敲辯論 [FF0C?] 各個黨派意見折衝後形成的條例和制度 [FF0C?] 但這不能否認在沒有文字洗禮的部落社會照樣有一套律法的規範 [FF0C?] 那就是自我組織的結果。

大腦本身也是自我組織形成的系統。大腦的演化或發展是自我組織 [FF0C?] 因為大腦的知覺認知功能依靠經驗學習而來 [FF0C?] 而支持知覺認知功能的大腦結構需要外在的感官刺激才能發展。基因的指令理論上不足以精準的決定神經元的各種連結。另外 [FF0C?] 每個人的腦都各有其特性 [FF0C?] 其發展路徑也無法預知。由於經驗的不同 [FF0C?] 神經元之間藉著軸突與樹枝狀的連結以及神經元突觸網絡 [FF0C?] 隨時產生變化 [FF0C?] 之後隨著年長 [FF0C?] 才逐漸定型。大腦的自我組織基本上有賴於感官知覺經驗的刺激 [FF0C?] 並進一步優化基因神經元連結所構成的基本藍圖。

在自然界 [FF0C?] 雁行隊伍 [FF0C?] 蜂窩或白蟻窩是經常被提到的例子。螞蟻分泌的費洛蒙是螞蟻溝通的主要手段 [FF0C?] 只能傳達大約十種左右的訊息 [FF1B?] 但集體的螞蟻卻能建造含有好幾百萬的超級蟻窩 [FF01?] 圖 1-1 白蟻窩壯觀的模樣 [FF0C?] 令人佩服其「自我組織」的奧妙 [FF0C?] 也是生物學家或電腦演算學者極感興趣的研究對象。讀者如想進一步了解白蟻如何分工合作 [FF0C?] 如何表現集體智慧 (swarm intelligence)[FF0C?] 白蟻窩內部的建築如何解決採光或通風的問題 [FF0C?] 可參見 (**Margonelli2012**)。

圖 1-1「自我組織」的白蟻窩

(http[FF1A?]//upload.wikimedia.org/wikipedia/commons/7/73termite_cathedral_DSC03570.jpg)

大城市的形成 [FF0C?] 例如台北市 [FF0C?] 也是一個複雜且具有適應性格的系統 [FF1B?] 因為它需要提供幾百萬的市民一切生活所需 [FF0C?] 但實際上台北市或任何城市的形成並沒有任何指揮中心做統籌的工作 [FF0C?] 而是來自許許多多市民的互動。有了互動 [FF0C?] 就有回饋 [FF0C?] 尤其是逆向回饋。如此長期循環調適 [FF0C?] 而造就了我們所熟悉的城市的基本面貌。城市都有人行道 [FF0C?] 而街道也不宜太長 [FF0C?] 最好容許商業住宅混用 [FF0C?] 以求增加人流物流的互動 [FF0C?] 也就是促進回饋 [FF01?] 在食物方面 [FF0C?] 一個城市有多少傳統市場 [FF0C?] 有超市 [FF0C?] 有小商店 [FF0C?] 有小吃店 [FF0C?] 有餐館 [FF0C?] 有路邊攤 [FF0C?] 貨物流動進出完全自由開放。餐館商家起起落落 [FF0C?] 也是適者生存 [FF0C?] 弱者淘汰的生存之道 [FF0C?] 但沒有任何市政單位管控商家起起落落這些事宜。城市的形成在這些方面有其自然演變的邏輯 [FF0C?] 靠的是市民的需求 [FF0C?] 他們有唇齒與共的生存目標: 有些人需要吃的穿的 [FF0C?] 有些人則提供吃的穿的 [FF0C?] 彼此共存共榮。城市隨著時間的推移 [FF0C?] 面貌容或有變化 [FF0C?] 如壽司店變得生意鼎盛 [FF0C?] 而影響附近的比薩店家 [FF0C?] 但這不影響整個大城市的基本性格 [FF0C?] 就是它是個複雜且具有調適能力的系統。

語言跟城市一樣是個複雜且具有調適能力的系統。複雜指的是一個語言社群的人口眾多 [FF0C?] 彼此互動頻繁 [FF0C?] 每個人有共同的社會目標 [FF0C?] 那就是希望能相互溝通 [FF0C?] 但每個人的認知習慣、知覺偏好、生活經驗、講話的格調各有不同 [FF0C?] 方言的形成 [FF0C?] 或語言的變化都跟這些因素有關。雖然如此 [FF0C?] 這樣的語言社群很自然地會從廣泛而龐雜的語言互動中「自我組織」進而衍生出一套系統、一套規律、一套語法。這種自我組織而產生的語法系統有學者稱之為呈現語法 [FF08?]emergent grammar[FF09?]。這是因為在社群裡的人有

共通的目標 [FF0C?] 需要相互溝通。每一個人都使用各種可能的方法以求達到溝通 [FF0C?] 基於人性共有的認知策略 [FF0C?] 相似的推理能力 [FF0C?] 社會化的約束 [FF0C?] 久而久之 [FF0C?] 就產生一個跨個人 [FF0C?] 跨情境共通的語言結構體。這就是語言的系統。語法的呈現是「自我組織」的結果 [FF0C?] 因為語法系統的形成絕對沒有任何中樞指揮居間操控。

呈現語法 [FF08?] 在此應該把「呈現」一詞理解為修飾語 [FF0C?] 而非動詞 [FF09?] 是認知功能語言學中最為重要的概念之一。呈現的觀念很接近英國社會學者 Giddens[FF08?] 紀登斯 [FF09?] 所提出的「結構化 [FF08?]structuration[FF09?]」的概念。結構在此指的是傳統、制度、典章、道德規範、以及其他種種有約束力的行為準則。從社會學的角度 [FF0C?] 結構 [FF08?] 或組織 [FF09?] 一方面制約每個人的行為 [FF0C?] 另一方面也使你我的行為成為可能 [FF0C?] 也提供解釋你我行為的基礎。在語言 [FF0C?] 情況基本上也一樣 [FF0C?] 甚至可能更為清楚。所謂結構在語言指的是語法。有了語法系統 [FF0C?] 講話就有所規範 [FF0C?] 某些講法就不被認可接受 [FF0C?] 但也由於有了語法系統 [FF0C?] 你我可以表達想要表達的任何意念。呈現語法不是一個固定不變的系統。說話者隨著跟他人長期互動經驗的累積 [FF0C?] 導致調適 [FF0C?] 進而 [FF08?] 但並非有意識地 [FF09?] 改變自己的語法系統 [FF0C?] 亦即語法系統在心智的表徵 [FF08?]mental representation[FF09?]。一般所謂「句型」指的其實就是指語法系統中那些比較穩定的結構。

語言變化與混沌現象

語言是個動態系統 [FF0C?] 經常變化 [FF0C?] 這是常識。但語言的變化一定是語言的社群對某一個初始條件有某種敏感度。語言的變化往往不可預測 [FF0C?] 因為要有能力預測變化一定要對導致變化的初始條件有相當程度的了解 [FF0C?]

問題是我們通常沒有能力事先了解初始條件。有時候 [FF0C?] 很微小的變動 [FF0C?] 最後很可能造成明顯的語言變化。因此語言的變化基本上是一種蝴蝶效應的表現。有些學者則視為混沌現象 [FF08?]chaos[FF09?][FF0C?] 指的是一個非線性的動態 [FF0C?] 一開始看似毫無章法的現象 [FF0C?] 最後卻變成社群所普遍接受的有規律性的結構。請注意「混沌」一詞並非指混亂失序 [FF0C?] 而是指任何能自然而然從一個非線性的動態系統產生的複雜而有序的結構現象。具有混沌現象的系統在發展過程中很容易受各種外在因素的波動 [FF0C?] 也由於這個原因 [FF0C?] 它的演變方向通常無法預測。無論是蝴蝶效應或混沌現象 [FF0C?] 都說明語言結構經常變動不居 [FF0C?] 只是乍看之下像是無序混亂的現象最後卻都能形成有序的結構體 (關於混沌現象請見科普著作 Gleick1988 ; Waldrop1992)。

### 0.1.4 語言既非自然物 **[FF0C?]** 也非人造物

在這裡應該稍微解釋語言變化的基本原理 [FF1A?] 語言跟社會現象或社會制度 [FF08?] 含法律、宗教、貨幣、市場經濟等 [FF09?] 一樣 [FF0C?] 都是集體行為自然產生的結果。但那個結果不是社群裡的行為參與者原來有意圖要達到的目的 [FF0C?] 而是諸多集體行為產生的副作用之一。語言會變化 [FF0C?] 但變化的結果不可能是你我一開始就有意想加以改變而使然。在一般情況下 [FF0C?] 你我個人既沒有那個能力 [FF0C?] 也沒有那個意圖。因此 [FF0C?] 語言變化中冥冥中似乎有一隻隱形的手在操弄 [FF0C?] 其實那只是集體行為下產生的意外效應。語言跟其他人文現象或社會制度一樣 [FF0C?] 是德國學者 (**Keller1994**) 所謂的第三類現象 [FF0C?] 有別於第一類現象 [FF0C?] 指的是自然界的自然萬象 [FF0C?] 這些基本上都獨立於人類的參與 [FF0C?] 也有別於第二類現象 [FF0C?] 指的是一切人造的器物 [FF0C?] 因此是人類意志的表現。但語言似兼具第一類跟第二類現象的特質 [FF0C?] 因為語言是人類在演化過程中創造的 [FF0C?] 這像是第二類現象 [FF1B?]

另外一方面 [FF0C?] 語言的演變人類無法左右 [FF0C?] 這很像是第一類現象。總之 [FF0C?] 語言的變化是許許多多個別的行為集體造成的 [FF0C?] 但是造成的變化是偶然的意外 [FF0C?] 因為沒有人有能力 [FF0C?] 或有任何意圖要創造那些變化。

我們可以用一個很恰當的比喻來理解這個中的道理。那就是「路是人走出來的」這句話。假定你住的新蓋公寓大樓距離一家銀行不遠 [FF0C?] 但是市政府只開了兩條馬路 [FF0C?] 一條是左轉出去 [FF0C?] 再接一條右轉的路到達銀行。其實最直接最方便的走法是走斜邊過去。果然公寓蓋完沒多久 [FF0C?] 住戶就走出這樣的一條斜路了。一開始這樣走的人並無意要開發新的路 [FF0C?] 他或她只是圖個方便 [FF0C?] 找一條最省時省力的路徑 [FF0C?] 但漸漸地其他住戶也有同感 [FF0C?] 有同樣的目標 [FF0C?] 就是找個同樣省時省力的走法。就憑著這許許多多的個別行為 [FF0C?] 最後不約而同走出一條從公寓到銀行有別於政府蓋的新馬路 [FF01?] 語言或社會制度絕大部分都是這樣的第三類現象。

### 0.1.5 投射與語言的時間性

要了解語言的結構 [FF0C?] 第一個要件是了解人與人在互動時所講的話語是在時間之流上逐次展開的音聲現象。在對話時 [FF0C?] 每個人都有話語權。我講完了 [FF0C?] 你接著講。話語權是協商的結果。也就是在聽話者默許 [FF0C?] 暗示或同意之下 [FF0C?] 講話者才有話語權。而且通常每個人是一句話一句話的講 [FF0C?] 每一句話約花 1.5 秒 [FF0C?] 可長些 [FF0C?] 也可短些。另外 [FF0C?] 由於大腦天生的預測性格 [FF0C?] 聽話者也有能力預知何時對方會結束話語 [FF0C?] 輪到自己有話語權 [FF0C?] 而轉而成為講話者。這個預知能力如果表現在對話或行為的預測方面 [FF0C?] 就稱為投射 (projection)。投射就是一種預測。聽話者利用他/她整體的語法知識 [FF0C?] 包含話語的聲調節奏以及對語言結構區塊

(chunking) 的掌握 [FF0C?] 投射/預測對方的話語的走向 [FF0C?] 從而接話、或插話、或幫對方完成話語。了解語言的時間性比較容易了解語法結構 [FF0C?] 乃至於語法系統到底如何呈現。

投射可以是指局部句子內的投射 [FF0C?] 也可以是指跨越句子的投射。講話者講到某一個語詞 [FF0C?] 但搜尋遇到困難 [FF0C?] 講不出那個詞 [FF0C?] 這時聽話者往往能藉著投射能力 [FF0C?] 預測到那個語詞 [FF0C?] 幫他/她講了 [FF0C?] 就是一種局部投射。反之 [FF0C?] 如果講話者先講了‘因為如何如何’[FF0C?] 聽話者就可以投射下一句應該是‘所以如何如何’。如果講話者講‘我是想說’[FF0C?] 那麼聽話者就能投射講話者大概還有話要講。在下面 F 跟 M 的中文對話中先是 M 在 263 行用了熟語 □ 問題是 □ [FF0C?] 但是沒有講完。「問- 」表示講了「問」馬上停頓 [FF0C?]「問題是＝」的「＝」表示拉長「是」的發音 [FF1B?]「[FF08?]TSK[FF09?]」表示 M 講話稍有困難 [FF0C?] 需要思考一下 [FF0C?] 而舌尖去碰牙齦的動作所發的聲音。因此在 265 行 F 投射 M 必然還有話說 [FF0C?] 而等了 3.1 秒才說話 [FF0C?] 接著 M 在 266 行果然說出她心中的話。

(1)

((Actor))

260 F: .. 你為甚麼都不請豪寧去看電影啊。

261 M: …(1.43) 我想啊。

262 …(0.85) 問-

263 .. 問題是=,

264 …(TSK)

265 F: …(3.1) 是甚麼。

266 M: …問題是不知道怎麼 -

267 .. 怎麼找他去。

話語的時間性加上話輪的接續性使得許多話之間環環相扣 [FF0C?] 前一個話可以預知下一句話 [FF1B?] 下一句話也可以預知再下一句。久而久之 [FF0C?] 有些話語的形式、位置、功能等往往變得很定型 [FF0C?] 即所謂結構化。話語一旦結構化 [FF0C?] 聽話者更容易理解 [FF0C?] 更容易投射對方的下一句話。語言中有相當多的結構化的話語 [FF0C?] 難怪人與人之間的對話常常可以很順暢 [FF01?] 下面是兩個好朋友 H 跟 L 的對話。

(2)

((Marriage))

446 L: .. 那個時候有一個--

447 .. 男%-.. 男士,

448 .. 他的條件不錯。

449 H: ..mhm.

450 L: .. 結果那個時候呢 [FF0C?]

451 .. 他就提出跟我 =

452 H: .. 在一起。

453 L: ex:key:0 提出跟我結婚。

454 H: mhm。

L 在 451 行時拉長了她的「 我 」一字 [FF0C?] 表示她短暫陷入思考 [FF0C?] 因此 H 在 453 行就做了投射 [FF0C?] 試著替她完成話語 [FF0C?] 雖然不是 100% L 想說的話 [FF0C?] 但八九不離十 [FF0C?] 至少結構上「 在一起 」跟「 結婚 」都是很合理合法的補語句。

(3)

是一對母女的對話。D 先是跟她媽媽敘說她工作上的表現 [FF0C?]D 的第二次說的話「 媽 =/ 」[FF08?]「 = 」表拉長音; / 表音高有點拉高 [FF09?] 再簡單不過了

[FF0C?] 但我們聽得懂她撒嬌的模樣。這完全歸功於語境的作用 ── 所謂語境就是考慮話語的形式、位置與功能 [FF0C?] 而推知「媽＝/」在 [FF08?]4[FF09?] 這樣的對話情境下說話者女兒撒嬌的模樣。

(4)

D: 我花了很多心力 [FF0C?] 才有這一點成果。\

M: 妳交男朋友有這麼賣力就好了。\

D: 媽＝。/

語言的投射能力是大腦預測性格的一環 [FF0C?] 也因此可以推想人類創造出來的事物都常常內建有這個預測性能。一個建物的大門入口如果有個把手模樣的東西 [FF0C?] 表示你要拉開門 [FF1B?] 反之 [FF0C?] 如果是個貼片模樣的東西貼在門上 [FF0C?] 那表示你要推開門進去。人造器物常常有這個內建的預測性能。

[FF08?]attractor[FF09?]]1.6 語言的複雜性 [FF1A?] 吸引子 [FF08?]attractor[FF09?]

上面提到語言是個複雜而具有調適性格的系統 [FF0C?] 指出一個語言社群很自然地會從廣泛而龐雜的語言互動中「自我組織」進而衍生出一套系統 [FF0C?] 一套規律或一套語法。這種自我組織產生的語法系統就是語法的呈現。但這裡語法的呈現既指每個人藉著長久與他人互動而形成的具有個人風格的語言 [FF0C?] 也指整個大的社群藉著長期彼此互動而形成的語言。無論是個人的語言 [FF0C?] 乃至社群的語言都是「自我組織」的結果。大的社群語言裡含有許許多多具個人特殊風格的小語言。社會語言學家的研究早就指出這個現象 [FF1A?] 社群語言中往往含有大量的有規律的個別差異 [FF01?] 有一部微電影叫作梨子的故事 [FF08?]Pear film[FF09?] 中文版的梨子的故事語料中就可以看出有許多這一類的個別差異現象 [FF1A?] 這部六分鐘的微電影中 [FF0C?] 一開始可以是有個老人在樹上摘水果 [FF0C?] 接著是一個人牽著一隻羊走過來 [FF0C?] 然後老人摘的水果被一個路過的小孩偷走 [FF0C?] 接著那個小孩又遭遇了一些事等等。我們請了 20 個

台灣的大學生來講這個故事, 包含這個第一幕 [FF0C?] 但只有八個人提到這個場景。他們的講法可以歸納為四種句型 [FF1A?]

ex:key:5 a. 有一個人牽了一條小羊走過這個路邊。1 You NP V-le MP

b. 有一個牧羊人牽了一頭羊。就走過了。1 You NP V-le MP

c. 有一個人牽著一頭羊…走過去。5 You NP V-zhe MPD

d. 然後就有一個人呢 [FF0C?] 就拉著那個小牛 [FF0C?] 就經過。1 You NP V-zhe P

(You ‘有’; NP : 名詞片語; V : 動詞, -le ‘了’; M(manner of motion) : 走;

P (path) : 過; D (deictic) : 去)

這四個句型中 (c) 最精簡 [FF0C?] 最語法化 [FF1B?](d) 最不語法化 [FF0C?] 也顯示說話者過度使用「就」一詞。介系詞來自動詞 [FF0C?] 因此用介系詞是比較語法化的句法。相較之下 [FF0C?]20 個講這個故事的美國學生中有 18 個提到這個場景 [FF0C?] 但其中有 11 個人講法都用下面 ex:key:6 的句型 [FF0C?] 六個人用 ex:key:7[FF0C?] 一個人用 ex:key:8 的句型。(6) 用介系詞片語 with a goat 表達牽了一隻羊的概念 [FF0C?] 比中文的 (5c ) MPD 更語法化一點 [FF1A?]

(5)

And a man **comes along with** a goat 11

(6)

…and a man **comes by leading** a goat 6

(7)

…and this man is **pulling** a goat. 1

(5c) 或 ex:key:6 是分別在中文 [FF0C?] 英文兩個不同的系統中算是比較穩定的句型 [FF0C?] 因為最多人使用此一句型 [FF1B?] 其他的句型或講法都是變異。推廣而言 [FF0C?] 任何場景的描敘 [FF0C?] 大部分的人通常都會趨向使用有某一種

句型 [FF0C?] 一種有些學者稱為吸引子 [FF08?]attractor[FF09?] 的句型 [FF0C?] 就是我所謂穩定的句型 [FF08?] 也是教科書最常教給學生的那些句型 [FF09?]。雖然如此 [FF0C?] 我們也幾乎可以確定一定有少數人使用其他不同的 [FF0C?] 比較不穩定的講法。這是任何複雜而有調適性格的系統必然的現象 [FF1A?] 有穩定的結構的一面 [FF0C?] 但也有變異之處。語言的動態平衡 [FF0C?] 指的是語言的結構是個動態的現象 [FF0C?] 有穩定之處 [FF0C?] 也有變異 [FF1B?] 不管是穩定或變異 [FF0C?] 都是在某一個時空環境下的產物。變異的結構可能不久之後反而成為穩定的結構 [FF1B?] 本來穩定的結構也隨著時間的變化成為微不足道的變異 (關於吸引子在中文句法或南島語語法的表現請參見第三章構式語法的討論以及 Huang2013, Huang2017)。

吸引子 [FF08?]attractor[FF09?] 不一定專指語言 [FF0C?] 也可以引申到任何有穩定狀態的事物。最簡單的例子是一個擺動的鞦韆 [FF0C?] 由於阻力作用 [FF0C?] 很快就停止擺動。即使有外力推了一把 [FF0C?] 但幾秒鐘之後就又回到靜止的狀態 [FF0C?] 因為這是鞦韆的引子狀態。又如一個社會新鮮人開始成為朝九晚六的上班族 [FF0C?] 很可能幾個月下來 [FF0C?] 她的作息 [FF0C?] 甚至通勤路線 [FF0C?] 假日休閒方式等等都可能趨於固定 [FF0C?] 達到近乎可以預測的地步。理論上她的日子可以過得很有變化 [FF0C?] 事實上不然。她上班的這種作息方式就是一種吸引子 [FF0C?] 一種穩定的狀態。假定幾年之後 [FF0C?] 她換了一個工作 [FF0C?] 在外商公司上班 [FF0C?] 經常出差國外 [FF0C?] 那又是另外一種作息方式 [FF0C?] 亦即她進入一個新的吸引子狀態。同樣 [FF0C?] 汽車已經問世將近百年 [FF0C?] 其外觀樣態層出不窮 [FF0C?] 專家估計世界上汽車已經超過 10 億輛 [FF0C?] 可以想見它外觀的多樣性 [FF01?] 但不論如何 [FF0C?] 汽車還是有它基本的樣貌可循 [FF0C?] 不可偏離。這個不可偏離的外觀就是它的吸引子狀態。有了穩定的吸引子狀態 [FF0C?] 就有變異之處。汽車就是這樣的器物 [FF0C?] 它可以

有著保守的外觀 [FF0C?] 也可以是弄得很炫。年度汽車展的賣點不外就是要靠新奇的外觀或內部擺設取勝。

### 0.1.6 1.7 語言的規律性

我們一直在強調語言是一套系統、一套規律 [FF0C?] 因此我們不能不談談規律性這個概念。音韻的結構有規律性 (見第六章) [FF0C?] 語法的結構也有規律性 [FF0C?] 因此連帶地意義結構也有某種規律性。在語言學界常以語法一詞泛指語言整體的規律性。語法的規律性是指一個語言社群裡的人 [FF0C?] 經過長久複雜的互動與調適而產生的有規則的系統。如果我們仔細觀察三到六歲學母語的小孩 [FF0C?] 或是初學中文的外籍人士就很容易體會到語法的規律性。下面一些句子是學了兩年中文的幾位外籍人士所說的中文 [FF1B?](10) 是其中一位寫的中文 [FF1A?]

(8)

從我的公寓到我的工作。

他不舒服說英文。

我走的快到學校來。

我想去散步一趟。

我剛才到了 [FF0C?] 不知道發生什麼了。

我買東西買的很窮。

我騎車避免石頭。

全朋友一起喝酒。

觀眾被告訴一個簡單的故事。

他畢業了大學。

(9)

在我看來我家是在台北市的最好的附近。我意見的原因是因為我家離我公司很近 [FF0C?] 所以我不必坐公共汽車 [FF0C?] 或騎機車。下班的時候我不著急坐客滿的公車。開車或騎機車常常也很危險的。在路上不遵守交通規則的駕駛很多 [FF0C?] 他們常常在紅燈不停車。

這些外籍人士所講的 ex:key:9 前面幾句如果改為 ex:key:11 的講法應該是較合乎一般中文語法的「 規律 」[FF1A?]

(10)

從我的公寓到我工作的地方 [FF0C?]

他說英文 [FF08?] 還 [FF09?] 不是很自然。

我很快走到學校來。

我想去散步一下。

我剛剛到 [FF0C?] 不知道。

但語法的規律如何確定 [FF1F?] 如何發現規律的內涵 [FF1F?] 這些問題可就不容易有簡單的答案了 [FF01?] 就算講華語的人也不一定同意我上面 ex:key:11 的說法 [FF08?] 見第三章功能語法學進一步的討論 [FF09?]。現在假定讀者同意 ex:key:11 的講法的確比較可以接受 [FF0C?] 那麼如何解釋 ex:key:9 跟 ex:key:11 究竟怎麼不同 [FF1F?] 也就是什麼樣的語言規律讓我們可以解釋 ex:key:9/ ex:key:11 的不同 [FF1F?] 有一派比較傳統的學者認為我們一旦學會了一個語言 [FF0C?] 我們就懂很多的詞彙 [FF0C?] 而且能講無窮的句子 [FF0C?] 並且能懂無窮的句子。因此語言的知識基本上就包含兩種成分 [FF1A?] 辭典與語法。辭典指的是一個語言所涵蓋的詞彙 [FF0C?] 而語法則指規律 [FF0C?] 即能把單字串成合乎語法的詞組 [FF0C?] 然後把合乎語法的詞組串成合乎語法的句子的那些規律。但所謂辭典不是手摸得到的那種辭典 [FF0C?] 語法規律也不一定指的是那些可以一一條列出

來 [FF0C?] 擺在眼前供人檢視的規則。語法學者認為 [FF1A?] 所謂辭典應該是指心理辭典 [FF0C?] 儲存在大腦的辭典 [FF0C?] 它可以產出許多但無法窮舉的詞彙。同樣的 [FF0C?] 所謂語法的規律指的是儲存在大腦的一套複雜的規則 [FF0C?] 讓每個人有能力講出無限多的句子 [FF0C?] 聽得懂無窮多的句子的那些規律 [FF0C?] 雖然在實際操作上卻還沒有人有能力把這一套規律一五一十的條列出來。其實真正的困難在於 □ 語法的規律 □ 如何確定根本無解 [FF0C?] 常人跟專家的看法也往往大相逕庭 [FF0C?] 因為「合乎語法」追根究柢是個困難複雜而且高度模糊的概念。

上面我們提到 ex:key:9 的那些句子不好 [FF0C?] 不合乎華語的語法。「合乎語法」在生成學派指的是純形式而毫不考慮意義的規律 [FF0C?] 但較常見的看法是指合乎以中文為母語的人認為是正確的可以接受的講法。問題是 [FF1A?] 如果我們仔細聽身邊的人講話 [FF0C?] 或電視上的大人物講話 [FF0C?] 並且記錄下來 [FF0C?] 慢慢品味 [FF0C?] 你會發現他們講的話其實很多是怪怪的 [FF0C?] 甚至根本不合乎你我熟悉的講法 [FF0C?] 也就是不合語法 [FF01?] 如果我們搜尋一下英語或中文的語料庫 [FF0C?] 也會發現其中有不少句子我們覺得很怪異。怎麼辦 [FF1F?] 事實上你我覺得不可以接受、不合乎語法的句子其他人很可能覺得可以接受 [FF0C?] 不會排斥 [FF0C?] 不然怎麼會被收進語料庫 [FF1F?] 當然啦 [FF0C?] 語料庫的句子 [FF0C?] 尤其是口語語料 [FF0C?] 是據實紀錄 [FF0C?] 因此很可能講話的人一時口誤 [FF0C?] 注意力不集中 [FF0C?] 完全沒有心防下的產出 [FF0C?] 使得那些句子不盡完美 [FF0C?] 講法或有改進的空間。這一類的情況說明單純希望藉著語料庫建立合乎語法的準則也有其盲點。一個比較中肯的看法是認為如果在一個社群裡 80% 的人同意某一個講法是正確的 [FF0C?] 可以接受的 [FF0C?] 那麼基本上那樣的句子就是合乎語法 [FF0C?] 或是說已經語法化 [FF0C?] 已經成為語法系統的一部分。假定大部分的讀者同意上面 ex:key:9 的句子確實不理想 [FF0C?]

但 ex:key:11 就沒有問題。這就合乎我們所謂 80% 的通則。

但我們還沒真正解釋何以 ex:key:9 的句子不理想, 而 ex:key:11 沒有問題。如果我們耐心坐下來 [FF0C?] 慢慢思考 [FF0C?] 我們會發現語法也許有幾個簡單的大的規則可言, 但其實例外更多。20 世紀中葉美國語言學家 [39]Sapir1921 有一句名言: "…Unfortunately, or luckily, no language is tyrannically consistent. All grammars leak." 由於「語法的規律」的界定困難重重 [FF0C?] 因此最近認知學派的學者主張語言的研究應該擺脫生成語法學派那種對「語法規律」不必要的甚至有點誤導的投入 [FF0C?] 轉而關心語法系統如何呈現 [FF08?]emergence[FF09?][FF0C?] 語言如何定型化 [FF0C?] 就是上一節提到的吸引子如何呈現。這是認知語言學派的主張 [FF0C?] 也是目前多數的語言學者比較認同的想法。

### **1.7.1** 詞意/概念的模糊性

「語法規律」迄今無法建立一套多數學者接受的準則。詞彙的意義基本上也同樣模糊。廣泛言之 [FF0C?] 人類所有的概念基本上都如此 [FF0C?] 不僅僅只限於跟語言有關的概念。詞彙意義模糊是指某一個單字的詞義的界線不清楚。「合乎語法」一詞就是很好的例子。這是語言的本質之一 [FF0C?] 但卻不能說是語言的□缺點□。如果兩個人某種原因一旦開始辯論 [FF0C?] 其中一方很快就會發現對方需要把自己的想法先定義清楚 [FF0C?] 才能釐清爭論的焦點所在。說話者常常需要自己先確定自己要表達的意念是甚麼 [FF0C?] 盡量把界線弄清楚一點 [FF0C?] 這就是因為一般人了解要把自己的話定義清楚的重要性。

何以說詞彙意義的模糊性不能算是語言的□缺點□[FF1F?] 因為詞彙意義之所以模糊來自於人類知識論上的侷限 [FF1A?] 人類感知能力的局限從而導致語言的模糊性。我們試以簡單的「高」一詞為例。「高」的詞意模糊 [FF0C?] 這很清楚 [FF0C?] 無須辯解。高矮是相對的 [FF0C?] 但我們有辦法使它不模糊嗎 [FF1F?] 不

可能。假定我們同意 160 公分是矮 [FF0C?]161 公分甚至 165 公分以上就不矮。但沒有人能單憑知覺能力判斷某人是 160 公分高 [FF0C?] 或 161[FF0C?] 甚至是 165 公分高。這是知識論上的問題 [FF1B?] 不單純是語言的問題。幸好我們現在有能力測量身高 [FF0C?] 解決一部分的爭議 [FF0C?] 因此最簡單的解決方法是保有高或矮這樣的詞彙 [FF0C?] 以供平常使用 [FF0C?] 但遇到爭端時 [FF0C?] 則先澄清立場 [FF0C?] 然後用非語言的輔助手段去解決爭論。很可惜 [FF0C?] 大部分情況下我們缺少這樣的輔助手段。

詞義概念固然模糊 [FF0C?] 實際生活層面上很多時候我們對概念與概念之間的界線也由於知識論上的困難缺乏該有的敏感度。假定許多人爭相逃難 [FF0C?] 擠到一艘船上去。船只能載重兩噸 [FF0C?] 或 80 人。但誰能保證或預知 81 人 [FF0C?] 甚至 85 人一定超載 [FF0C?] 而忍心把第 86 號那個人趕下船 [FF0C?] 讓他/她絕對上不了船 [FF1F?] 沒有人。但很有可能第 86 人一上船 [FF0C?] 真的超過負荷了 [FF0C?] 船就沉了 [FF01?] 我們搭電梯通常就沒有這個問題。一旦超載 [FF0C?] 電梯就發警鈴 [FF01?] 如果所有的器物或制度都有內建的預警機制最為上策 [FF08?] 雖然這還是不能否認「詞義是模糊的」這個知識論上的難題 [FF0C?] 否則也不需要有任何的內建預警機制 [FF09?]。

### 0.1.7 語言的社會性 [FF1A?] 共識 [FF0C?] 背景條件與推論

語言是一個社群經過長久的互動產生的複雜而有調適性格的系統 [FF0C?] 因此它的社會性是毫無疑問的。在一個社群裡大家認同彼此講的是同一個語言 [FF0C?] 也遵循使用這個語言的基本社會規範 [FF0C?] 例如講話時的話語權 [FF1B?] 也體認到在使用語言時需要雙方互有共識或默契。有趣的是只要兩個人在講話 [FF0C?] 聽話者經常要做一些推論 [FF0C?] 才能溝通無礙。因為溝通幾乎一定是建立在某些基本的共識或是默契之上。下面夫妻間的對話就是個例子

[FF1A?]

(11)

先生 [FF08?] 人在樓上 [FF09?]: 我那個那個放在哪?

太太 [FF08?] 人在樓下 [FF09?]: 在抽屜裡。

在這個對話中男的甚至沒有講出他要找的東西 [FF0C?] 而太太居然能毫不費力的就知道答案 [FF01?] 這是歸功於兩人之間的默契與共識。大概先生即將出門 [FF0C?] 在找某樣出門前常帶的東西, 太太了然於懷。同樣 [FF0C?] 假定你跟朋友分別被告知說下個禮拜一在台北碰面 [FF0C?] 到了那一天 [FF0C?] 彼此都到台北火車站碰面 [FF0C?] 雖然事先沒有被告知說要在哪裡見面。這是因為大家有了某種共識或默契。火車站往往是交通最便捷最容易到達 [FF0C?] 大家最常見面約會的地點 (相形之下 [FF0C?] 在人煙稀少地處偏遠的小鎮 [FF0C?] 郵局反而可能是大家約會的場所)。這種基於共識而獲得一致的‘問題解決’方法是諾貝爾獎得主經濟學家 Thomas (**Schelling1960**) 的協調競合遊戲中所謂的焦點 [FF08?]focal point[FF09?][FF0C?] 指的是那個彼此在有共識下獲致的解決方法。一旦有了共識或默契 [FF0C?] 你我的對話就順暢得多!

我們說話時 [FF0C?] 使用的語言有上面提到的那種模糊性 [FF0C?] 但我們心中的意念往往比較精準。‘她長得很可愛’一句話是有點模糊 [FF0C?] 聽話者一定不確定你所謂「可愛」是甚麼模樣 [FF0C?] 但說話者心目中她的可愛的樣子一定很清晰。這就涉及到語用學所謂的「說話者的意念」[FF08?]speaker meaning[FF09?] 這個概念。一個句子的語義跟說話者的意念 [FF08?] 說話者想表達的意義 [FF09?] 可以是兩回事。它們之間幾乎永遠都有個鴻溝 [FF1B?] 這個鴻溝就是藉著推論來跨越。意思是說我講的話語常常只講個大要而已 [FF0C?] 本身並無法真正完全表達我心中的意念 [FF1B?] 是聽話者藉著推論才能理解你的話。有時候說話者講的是隱喻講法 [FF0C?] 或是誇大講法 [FF0C?] 或是很間接的言語行為 [FF0C?] 這些都

要聽話者藉著推論才能理解說話者的本意。

有些推論嚴格講是雙方都預設必須接受的基本的背景要件。例如

(12)

a. 老師在黑板上寫字。

b. 老師在講台上寫字。

c. The bus is at the door.

d. The mailman is at the door.

前面兩句除了介系詞片語不同之外 [FF0C?] 其他沒有不同 [FF0C?] 但兩個句子的了解完全不同 [FF1A?](a) 老師人不在黑板上 [FF0C?](b) 則老師人當然是在講台上。(c)、(d) 兩句 [FF0C?] 巴士跟大門間的距離以及郵差跟大門的距離很不同 [FF1A?](c) 通常是指巴士就停在路邊 [FF08?] 也許對著大門 [FF09?][FF0C?] 但 (d) 郵差應該就站在大門前才對。前面兩句用了不同的介系詞片語 [FF0C?] 因而預設了在教室情境下衍生的不同的背景知識 [FF1B?] 也就是雙方認為是必要的共同的理解 [FF0C?] 雖然句子沒有明言 [FF0C?] 卻一定要成立 [FF0C?] 才滿足這句話的「說話者的意念」。這個背景知識也是說話者所要表達的意念的一部分。因此雙方溝通時要得到合情合理的‘解決方法’就是接受預設背景知識成立的事實而無須明言。

語言 **(**的使用**)** 是一種素描 **[FF1A?]** 客體與背景

上面 ex:key:12 先生與太太之間的對話簡短 [FF0C?] 但不影響溝通。其實人與人的溝通講常常只講求大要而已 [FF0C?] 是聽話者藉著推論才能理解你的話。由於語言的社會性 [FF0C?] 意義的模糊性以及人類知覺認知系統的不對稱性 [FF0C?] 說話者往往習慣性的只勾勒一個很粗淺的素描 [FF0C?] 而期盼聽話者能自動填補話語不足之處 [FF0C?] 就像畫石膏像的素描一樣 [FF0C?] 很自然地會略去許多細

節。這是常態。講話比較像是畫素描。如果你反其道而行 [FF0C?] 試著過度描繪細節 [FF0C?] 結果有如法律條文一般拗口 [FF0C?] 反而失去常人語言該有的自然與樸質。

了解語言的使用何以經常像是畫素描一樣 [FF0C?] 一個方法是看看你我講話時常常省略甚麼成分 [FF0C?] 保留甚麼成分。上面 ex:key:13a 的句子 [FF08?] 老師在黑板上寫字 [FF09?] 表示老師這個人是在教室 [FF0C?] 但是不在黑板上 [FF0C?] 在黑板上的是老師寫的那些字。那為甚麼不明白講出她人是在教室 [FF1F?] 這其實是反映人類的知覺/認知系統常見的不對稱現象。你看到老師在教室 [FF0C?] 看到她在黑板上寫字 [FF0C?] 你的視覺系統很自然地讓你看到寫字的老師是客體 [FF0C?] 教室是背景。在我們的認知系統中 [FF0C?] 客體通常是指行動或移動中的人或物 [FF0C?] 因此最容易被知覺系統所察覺 [FF1B?] 背景則通常是靜態 [FF0C?] 因此是最容易被忽略的成分。這是知覺系統的不對稱性。你的認知系統也就自然地想像有個客體從背景中凸顯出來。一旦你把認知系統化為語言時 [FF0C?] 就自然地用了類似 ex:key:13a 那樣的句子 [FF1A?] 寫字的老師是客體 [FF0C?] 教室是背景。背景不是我們所感興趣的對象。

因此我們講話時之所以經常略去某些話語不講 [FF0C?] 不單純是語言的現象 [FF0C?] 而是基本上來自人類的知覺/認知系統的不對稱性 [FF0C?] 進而表現在語言結構上客體/背景的不對稱性。下面幾個句子同樣很能說明這種不對稱性。

(13)

a. 他昨天到了。

b. 你去倒垃圾。

c. 飛機滑行了一下就起飛了。

d. 我勸不動他。

(14a) 省略了他到達的地點。用移動動詞表達移動時 [FF0C?] 移動者是客體

[FF0C?] 移動的路徑或目的地是背景 [FF0C?] 因此略去不提是很自然的。(14b) 的「倒垃圾」也是移動動詞 [FF0C?] 但是不像 (14a) 的「到」那種自行移動的動詞 [FF0C?] 而是有外力介入的使役動詞。「倒垃圾」的客體是垃圾 [FF0C?] 而「倒」的目的地 (例如垃圾桶) 是背景 [FF0C?] 因此在此也被省略了。(14c) 的「滑行」也是移動動詞 [FF0C?]「滑行」的客體是飛機 [FF0C?] 跑道是背景 [FF0C?] 也同樣略去背景不提。(14d) 的「勸」是言語動詞 [FF0C?]「勸」的行為涉及一個勸說者 [FF0C?] 一個被勸的人 [FF0C?] 以及勸說的內容 (例如要他放棄參加聚會) 。「勸」其實很容易被理解為一種類似使役移動動詞 [FF0C?] 勸說者把勸說的內容試著移動到被勸的人心中 [FF0C?] 說服他。在這一句 [FF0C?] 被省略的是勸說的內容 [FF0C?] 也就是移動的目的地 (有如把被勸說的人移動到某一目的地一般 [FF1B?] 一旦被勸的人移動到那個目的地 [FF0C?] 就等於他被勸動了) [FF0C?] 因此也同樣可以省略。

「經常省略」不等於「必須」省略。上面的那些句子確實都可以把省略的成分放回去 [FF0C?] 只是那就不是平常你我習慣的講法。一個社群自然發展出來的語言系統必有其生存之道 [FF0C?] 因此試著去研究何以某些成分可以省略 [FF0C?] 某些成分不可以省略 [FF0C?] 是過去四十多年語法的研究中一直持續吸引學者關注的議題 (參見 David2016)。

### 0.1.8 語言 [FF0C?] 推論與互動引擎 [FF1A?] 大腦是個喜愛預測的器官

語言的社會性表現在人類天生擁有一些社會互動的特性 [FF0C?] 即人類在對話的情境下有許多共識或默契 [FF0C?] 讓聽話者可以輕易地作出有效的推論 [FF0C?] 而說話者通常也預期聽話者有能力做正確的推論。大部分情況下雙方的預期也都可以得到滿足。一般而言 [FF0C?] 彼此互動越多 [FF0C?] 累積的共識或默契越深

厚 [FF0C?] 推論也越順利。有了共識 [FF0C?] 彼此可以心照不宣 [FF0C?] 因而彼此講的話語也可以更為簡略 [FF0C?] 正如上面那對夫妻的對話所顯示的。

由於推論的廣泛存在 [FF0C?] 可以節省說話者在產製句子時所需花費的「認知代價」(cognitive cost )。我們講話時速度並不很快 [FF0C?] 約 1.5 秒講 6 個音節 [FF0C?] 約相當中文四個詞 [FF0C?] 這比純粹的心理歷程 [FF08?] 指思想或推論 [FF09?] 慢上三、四倍左右。解決這個瓶頸的方法就是利用聽話者的推論來克服。在這裡「推論」指的是聽話者解讀對方話語的那種能力幸。好大腦是個喜愛並擅長預測的器官 [FF0C?] 隨時隨地都在預測你的下一步 [FF0C?] 對於你說的話 [FF0C?] 大腦也一樣 [FF0C?] 隨時隨地在推論你的下一句話。腦神經學者相信兩個人在對話時 [FF0C?] 彼此大腦活動在時空上亦步亦趨 [FF0C?] 相互耦合 (coupling)[FF0C?] 有時候說話者的某些腦區活動早於聽話者 [FF0C?] 但有時候聽話者的某些腦區活動 [FF0C?] 由於推測能力的作用 [FF0C?] 反而更早於說話者 (見 StephensEtAl2010) ! 這個耦合的能力越強 [FF0C?] 彼此的溝通越順暢。這裡所講的推論近乎諾貝爾獎得主經濟學家 (**Kahneman2011**) 所謂的那種比較快速的、直覺的 system 1 的思想 [FF0C?] 有別於比較花心思的而且有意識的 system 2 的思考 [FF08?] 參見第 12 章關於預測性的語言處理 [FF09?]。

人類擁有的社會互動性格絕對不見於其他動物。這些特性學者稱之為互動引擎 (EnfieldLevinson2006; Pagel2012 也有類似的觀點)。這個互動引擎表現在人與人之間的對話機制有其章法可循 [FF0C?] 知道何時啟動對話 [FF0C?] 何時可以插話 [FF0C?] 何時可以結束對話。其二是人類在對話時是高度合作的動物 [FF0C?] 知道一旦進入對話的情境 [FF0C?] 就是要遵循對話的機制 [FF0C?] 試著去了解對方的認知 [FF0C?] 對方的意圖 [FF0C?] 這就是合作。其三是人的互動有相互主觀性 [FF0C?] 我有溝通的意念 [FF0C?] 你也有溝通的意念 [FF0C?] 而且你我彼此了解雙方的心理或意圖 [FF0C?] 包括你的想法、意念、知識狀態等 [FF0C?] 然後根據這

些瞭解 [FF0C?] 預測行為與事件。根據心理學家的研究 [FF0C?] 九個月大的嬰兒就有解讀成人的意圖的能力 [FF0C?] 他/她會開始關注成人注意哪些事物。當成人專注看著嬰兒旁邊的杯子 [FF0C?] 嬰兒也會隨著成人的專注點去注意 [FF1B?] 她也會用手指指向某一個方向 [FF0C?] 告知大人正在找的東西的方位 [FF08?] 參見第 10 章兒童語言習得 [FF09?]。這說明小孩不但了解大人的心理狀態 [FF0C?] 而且知道自己的動作 [FF08?] 以手指指向某一方向 [FF09?] 可以被大人理解為具有溝通作用。自閉症患者就比較缺乏這種理解對方意圖或心智狀態的能力。

互動引擎的作用也表現在常見的溝通現象。假設你朋友走在你前面一個斜坡上 [FF0C?] 突然失去重心 [FF0C?] 摔了一跤 [FF0C?] 但隨即站了起來 [FF0C?] 轉身向你揮手致意。這時你知道你朋友沒事了。他揮手的用意是「希望」你「知道」他的意圖是要藉著揮手「表示」他沒事 [FF0C?] 而且他也「希望」你「了解」他的意圖。他「相信」自己沒事 [FF0C?] 這是第一層的心智狀態 [FF1B?] 我「知道」/「相信」他「相信」自己沒事。這是第二層的心智狀態。‘他的「用意」是要你「知道」他的意圖是藉著揮手表示她沒事 [FF0C?] 則是第三層的心智狀態。上面放在「」裡的動詞都是心理狀態或認知動詞 [FF1B?] 可見一個簡單的揮手的動作可以隱含複雜的心智活動 [FF0C?] 包含推論 (關於大腦的預測力請參見第 12 章第五節的討論)。

### 0.1.9 1.10 語言學簡史

語言學成為一門獨立的學問是很晚近的事 [FF0C?] 而這個新學門的誕生瑞士籍學者索緒爾 [FF08?]Ferdinand de Saussure[FF09?] 的貢獻厥功至偉 [FF0C?] 一般認為可以追溯自他 1916 年出版的法文著作 Cours de linguistique générale[FF08?] 普通語言學課程 [FF09?][FF0C?] 因此語言學作為一門學科而言 [FF0C?] 其歷史不過 100 多年而已。索緒爾這本書帶動了 20 世紀結構語言學派的興起

[FF0C?] 而他對語言符號的看法對後起的符號學的發展也扮演著先驅者的腳色。索緒爾認為語言構成一個結構體 [FF0C?] 組織意念 [FF08?]pensée[FF09?] 與音聲 [FF08?]*matière phonique*, *sons*[FF09?] 的結構體 [FF08?] 索緒爾用的是法文的 système 系統 [FF09?][FF0C?] 語言的核心性質在於其結構上的性質 [FF1B?] 結構造就了個別的成分以及個別成分之間的關係。結構的存在先於個別的結構的成分 [FF0C?] 意即語言的各個結構成分不是各自獨立 [FF0C?] 而是有賴於彼此相互界定其形式上的關係而存在 [FF0C?] 這個相互界定而產生的關係索緒爾在書中稱為符號的價位。例如先有下棋的整體結構的概念 [FF0C?] 然後個別棋子的存在才有意義 [FF1B?] 沒有下棋這個遊戲 [FF0C?] 就沒有棋子。我們觀看一盤棋 [FF0C?] 要看各個棋子在甚麼情況下擺在甚麼位置 [FF0C?] 而跟棋子本身的角色無甚關聯。日後學者就把索緒爾帶動的研究思潮稱為結構學派。索緒爾對結構語言學的貢獻在於他把語言區分成兩層結構 [FF1A?] 一層是比較抽象的無法直接觀察到的層次 [FF0C?] 稱為語言 (*langue*)[FF1B?] 另外一層是實際體現 *langue* 而成為我們可以直接觀察得到的話語層次 [FF0C?] 稱為話語 (parole)。在法文 langue 等於英語 language[FF1B?] 法文的 langage 則涵蓋 langue 跟 parole。語言學的研究主要是研究 langue 的結構 [FF0C?] 不是 parole。雙層結構的概念促成爾後聲韻學理論中音位/語音的區分以及語法理論中深層結構/表層結構的區分。

索緒爾把語言學當作是他所追求的廣泛的符號學的一支。語言學跟符號學一樣 [FF0C?] 語言的符號是聲音與概念之間的二元結構關係。任何符號一定包含兩個成分 [FF1A?] 符號形式本身 [FF0C?] 稱為意符 (或稱能指)[FF08?]Signifier[FF09?] 以及意符所指的意念 [FF0C?] 稱為意指 (或稱所指)[FF08?]Signified[FF09?]。在語言學發展的初期 [FF0C?] 符號多半指很簡單的詞 [FF0C?] 但在當代認知語言學符號也指複雜的構式。至於符號學則是研究文化社會中各式各樣的符號 [FF08?] 語言的以及非語言的 [FF09?] 的使用以及人類如何賦予這些符號意義。索緒爾認

為語言的符號 [FF0C?] 從聲韻的角度看 [FF0C?] 是任意而武斷的 [FF1B?] 從此符號的任意性 [FF08?]l'arbitraire du signe[FF09?] 廣為人知 [FF1B?] 但是另外一方面 [FF0C?] 從功能的角度 [FF0C?] 語言也是個表達意義而且有結構的系統。這個意義是人類互動後賦予的 [FF0C?] 是約定俗成的結果。

結構學派很快席捲了整個歐美人文社會學界 [FF0C?] 成為 20 世紀中葉最具影響力的一股思潮。在美國結構語言學的發展先後有幾個流派 [FF0C?] 其中一派深受當時心理學行為學派的影響 [FF0C?] 因此發展出一套比較側重形式分析而輕忽意義的學派 [FF0C?] 其代表人物當推布倫菲爾德 [FF08?]Leonard Bloomfield (1887-1949)[FF09?]。另外一派則比較注重語言在文化層面的意義 [FF0C?] 其代表人物是薩皮爾 [FF08?]Edward Sapir (1884-1939)[FF09?]。兩人都是廣義的美國結構學派的健將 [FF0C?] 還曾經在芝加哥大學同事過 [FF0C?] 也都分別著有一本深具影響力而同樣名為 Language 的書。布倫菲爾德的書出版於 1933 年 [FF0C?] 薩皮爾的書出版於 1921 年。有趣的是兩本書各發明了一句話以闡釋語法分析的要點。布倫菲爾德的書是用 Poor John ran away[FF1B?] 薩皮爾的書是用 The farmer kills the duckling。這兩位學者個性迥異 [FF0C?] 對語言的看法懸殊。布倫菲爾德致力於將語言學提升成為一門科學 [FF0C?] 因此排斥某些早期玄想哲學家的看法 [FF0C?] 以為語言能反映人類思考的普遍原理 [FF0C?] 也排斥那些試圖把邏輯方法套用到語言分析 [FF0C?] 而忽略語言的實際 [FF0C?] 更極力反對將拉丁文的語法系統奉為圭臬。布倫菲爾德早年留學德國 [FF0C?] 受教於當時最耀眼的新語法學派學者 Karl Brugmann[FF0C?] 研究過多種語言 [FF0C?] 包括德語、梵文 [FF0C?] 北美洲印第安語 [FF0C?] 也研究過菲律賓的官方語言塔加羅語 [FF0C?] 並撰成專書。布氏治學深受梵文專家 Panini 的影響 [FF0C?] 文章嚴肅冷靜 [FF0C?] 講究證據與方法論。布氏與其他幾位同好並於 1924 年創立了美國語言學會。薩皮爾才氣縱橫 [FF0C?] 研究範圍跨及語言學、人類學、心理學、文化研究、語言型態學等

多個領域 [FF0C?] 尤其是對北美洲印第安語言文化的研究與分類著力更深 [FF0C?] 對後代最具影響力。薩皮爾及其學生沃爾夫 [FF08?]Benjamin Lee Whorf (1897–1941)[FF09?] 提倡的所謂「語言相對論」[FF08?]linguistic relativity[FF09?] 更是一般知識份子耳熟能詳的假說。薩皮爾著作文采洋溢 [FF0C?] 讀之饒富興味與啟發性。前中研院院士李方桂當年就是薩皮爾在芝加哥大學任教時的第一個學生。

深受布倫菲爾德影響的結構學派其實有兩個盲點 [FF1A?] 其一是忽視意義對語言結構分析或解釋的重要性 [FF1B?] 其二是過度重視語言的表象 [FF0C?] 而忽視語言底層的結構或語言共通性的追求。第一個盲點導致 60 年代至 70 年代之間深受歡迎的生成語意學 (generative semantics)[FF0C?] 乃至 80 年代認知學派的興起 [FF1B?] 第二個盲點催化了 20 世紀 60 年代舉世聞名的杭士基 [FF08?]Chomsky[FF09?] 的變換語法 [FF08?] 後稱生成語法 [FF09?] 的誕生。杭士基區分語言的深層結構與表層結構 [FF0C?] 認為世界上的語言在深層結構方面基本上沒有不同 [FF0C?] 因為都來自同一個獨立的語言模組 [FF0C?] 而且所有的小孩天生就擁有涵蘊普遍語法的語言器官 [FF08?]language organ[FF09?]。小孩在語言習得過程中藉著調整普遍語法的幾個參數的不同可以很快習得母語。杭士基認為語法的研究目的在追求這個普遍語法。杭士基主導的生成語法有別於後起的生成語意學或之後衍生的認知語言學。杭士基認為生成語法追求普遍語法等於是對心智結構的探索 [FF0C?] 因此基本上是認知科學的一支。生成語法學派很快起了內部的理論爭執 [FF0C?] 其中一派認為深層結構必須建立在意義的研究的基礎上 [FF0C?] 不可能建立在純語言形式的/純語法的追求 [FF0C?] 因為語言的外在結構差異太大。

杭士基的涵蘊普遍語法的「語言器官」假說是最近 30 年來學界攻守雙方頗具爭議的議題之一 [FF0C?] 也被不少學者認為那最多只是個比喻 [FF0C?] 缺乏實質經驗內容 (「普遍語法」到底是甚麼? 爭論雙方又如何攻守? 參見

EvansLevinson2009)[FF0C?] 加上他的語言模組論又排斥語言結構跟一般的認知能力有任何關連 [FF0C?] 於是 70 年代中期起 (**Fillmore1976**) 首先提出 Frame semantics (框架語意學) 的主張 [FF0C?] 到了末期遂有認知語言學派的興起 (見 Behme2014; 關於框架語意學的討論見第四章)。認知語言學者 [FF08?] 有些學者偏愛“認知功能語法學派”一詞 [FF09?] 否認語言是個獨立模組的論點 [FF0C?] 強調訴諸認知心理學的概念 [FF0C?] 如記憶、知覺、分類、意義等對理解或解釋語言結構的重要性 [FF0C?] 因為處裡語言的認知能力跟處理其他現象的認知能力沒有不同。英國學者 Halliday 的系統功能語法 (systemic functional grammar) 也指出語言的研究不應該偏執於純形式句法的探討 [FF0C?] 而應該把語言理解為一種用來表達人類豐富的意念 (sense-making) 的資源。這是很有智慧的論點 (見 HallidayMatthiessen2004; 人類學者 (**Geertz1983**) 也認為文化基本上是深具符號意義的活動)。人文的世界是充滿意義活動 (semiosis) 的世界 [FF0C?] 而語言是這個世界與意義之間的界面。美國學者郎那克 (Langacker) 下面的觀點很有代表性 [FF1A?]

Meaning is what language is all about; the analyst who ignores it to concentrate solely on matters of form severely impoverishes the natural and necessary subject matter of the discipline and ultimately distorts the character of the phenomena described [12]Langacker1987

(見本書第 12 章提到的神經科學研究的重要成果 [FF0C?] 即 N400 或 P600 的現象 [FF0C?] 對認知學派的觀點是相當有力的佐證)。

認知功能語言學派中最具影響力的學者之一是剛剛提到的美國學者郎那克 (Langacker)[FF0C?] 而一般也把 1987 視為這個新派典的起點。因為這一年出版了三本認知語言學界重要的著作 [FF1A?] 郎那克的 Foundations of cognitive grammar、馬克. 強生 [FF08?]Mark Johnson[FF09?] 的 The body in the mind 、雷克

復 [FF08?]George Lakoff[FF09?] 的 Women, fire and dangerous things。1989 年國際認知語言學會宣告成立 [FF1B?]1990 年該學會的期刊認知語言學 [FF08?]Cognitive linguistics[FF09?] 也問世。成立學會以及出版代表學會的學術期刊都是一個學門走向制度化重要的指標。

綜合而言 [FF0C?] 過去 100 年語言學的四大派典的更迭可以整理如下 [FF1A?]

- 結構學派 [FF1A?] 索緒爾 [FF08?] 普通語言學課程 [FF09?]1916
  - ＊ 區分抽象的社區共有的符號系統 [FF0C?] 稱為 *language* [FF0C?] 以別於實際使用的話語 *parole[FF1B?]* 從知識論的角度 [FF0C?] 前者優先於後者。
- 美國結構語言學派 [FF1A?] 布倫菲爾德 [FF08?]*Language* 語言 [FF09?]1933
  - ＊ 發展語言學為講究證據與方法學的科學 [FF0C?] 惟較忽視意義或認知的解釋力。
- 生成語法學派 [FF1A?] 杭士基 [FF08?]*Aspects of the theory of syntax*[FF09?]1965
  - ＊ 提倡天生本領的語言器官說 [FF1B?] 鼓吹普遍語法的研究 [FF1B?] 側重抽象的語言形式而排斥一般認知功能的相關性。
- 認知功能語言學派 [FF1A?] 郎那克 [FF08?]*Foundations of cognitive grammar*[FF09?]1987

- 強調語意研究為一切語言研究之本 [FF1B?] 訴諸認知心理學的概念如記憶知覺、分類等理解或解釋語言結構重要性 [FF1B?] 正視運用為本的語料 [FF1B?] 堅信處裡語言的認知能力跟處裡其他現象的認知能力沒有不同。

有一點值得注意的是過去三、四十年認知功能學派最重要的論著也明顯地都有著跨語言型態學的取向。例如下列幾篇 [FF1A?](**Givon1979**)[FF1B?](**HopperThompson1980**)[FF1B?]Du (**Bois1987**); Talmy[FF08?]2000[FF09?][FF1B?](**Bybee2010**)。這樣的發展很自然 [FF0C?] 如果學界為了追求語言的普遍性 [FF0C?] 必須誠實面對世界上任何語言 [FF0C?] 也就是面對語言的多元性與歧異性 [FF0C?] 同時正視運用為本的語料與語料庫研究 [FF08?]Langacker1987:3[FF09?]。比利時學者 (**Geeraerts2010**) 指出語言學界的學風在過去三十年逐漸從生成學派轉向認知功能學派。這可從下表所列整個語言學界在過去三十多年學術界發表的論文理論取向及論文數量看出趨勢 [FF1B?] 尤其是從 21 世紀一開始這個派典轉向的態勢更為明顯 [FF1A?]

表 1-2 過去三十多年語言學界發表的論文之理論取向

| | 1988-92 | 1993-97 | 1998-02 | 2003-07 |
|---|---|---|---|---|
| 生成語法 | 304 | 538 | 337 | 296 |
| 認知功能語言學 | 81 | 337 | 376 | 916 |

有趣的是: 遲至 1989 年美國語言學會權威期刊 *Language* 的主編竟還如此宣示 [FF1A?] ‘an article in one of the central areas of our field which does not make reference to relevant generative research is unlikely to be accepted’(*Language* 65.2[FF1A?]445[FF09?][FF08?] 如果你投稿的論文不參考生成語法學派的論點 [FF0C?] 恐難被接受刊登 [FF09?]。

### 0.1.10 語言學的次領域

語言學基本的領域是語言本身結構的分析與解釋 [FF0C?] 諸如語音學、聲韻學、形態學語意學、語法學、語用學、歷史語言學、南島語言學、手語語言學等等。語言學研究的觸角廣及人文社會以及認知 [FF0C?] 生命或資訊科學有關的很

多領域 [FF0C?] 因此過去幾十年已經發展出不少次領域的學問 [FF1A?] 兒童語言習得 (見第 10 章)、語言與老化 (見第 14 章)、語言與手勢 (見第 11 章)、神經語言學 (見第 12 章)、社會語言學、語料庫與計算語言學 (見第 15 章、16 章)、生態語言學 [FF08?] 語言跟重要的生態議題的關係 [FF0C?] 生物多樣性的流失、生態的正義 [FF09?]、法庭語言學 [FF08?] 研究語言使用時所顯示的語音、詞彙、修辭、乃至語法方面的奇特之處以供法庭上民事刑事案件的審判時的證據 [FF09?]、應用語言學 [FF08?] 含語言教學理論與實際 [FF09?]、符號學、人類語言學、生物語言學 [FF08?] 研究語言生物基礎以及語言的演化 [FF09?]、臨床語言學 [FF08?] 研究語言病理 [FF0C?] 語言治療 [FF0C?] 如何矯正語言障礙以及吞嚥問題 [FF09?]、演化語言學 [FF08?] 研究語言能力在人類演化過程中如何出現、人類的遷徙與文化的演變、以及語言如何擴散到全球各個角落 [FF09?] 等等。很顯然的 [FF0C?] 由於篇幅所限 [FF0C?] 本書無法探討大部分的次領域以及科際領域。那應該是百科全書的使命。

### 0.1.11 語言學在台灣

語言學這門學問全面正式引進台灣是非常晚近的事。1969 年教育部才開始規定在大學部凡是語文相關的科系 [FF08?] 中文系、日語系、外語系等 [FF09?] 必須開授語言學導論的課程為必修課 [FF0C?] 雖然在此之前台大人類學系 [FF08?] 時稱考古學系 [FF09?] 已經將之列為必修課 [FF0C?] 由董同龢老師授課。估計當時台灣擁有語言學博士學位的不會超過 10 位。這個時候算是台灣語言學的萌芽期。不過有了這樣的必修課自然就吸引有興趣也有天分的學生進入這個領域。過了一個世代多的歲月 [FF0C?] 好幾個大學相繼成立了語言學研究所 [FF0C?] 到了 1995 年國內擁有廣義的語言學 [FF08?] 含語言教育學博士學位以及各個次領域 [FF09?] 的人數已超過 100 位 [FF0C?] 因此國科會 (現為科技部) 在 1995 年把語言學列為獨立學門

[FF0C?] 從此語言學的研究站穩了腳步。為了結合這股研究能量 [FF0C?] 國內幾位熱心學者在 1997 年發起成立台灣語言學會的籌備工作 [FF0C?] 並在第二年正式成立學會。時間上比起美國的語言學會整整晚了 74 年。

目前國內語言學的研究人力如何 [FF1F?]2006 年 ~2015 年前後十年間國科會 [FF08?] 即現今的科技部前身 [FF09?] 通過的研究計畫有 2867 件 [FF0C?] 平均每一年將近 300 件 [FF08?] 徐嘉慧等編著 2017[FF09?]。這 2867 計畫案中一般語言學佔 42%[FF0C?] 心理語言學與神經語言學佔 9%[FF1A?] 語料庫語言學與計算語言學佔 4%[FF1B?] 翻譯研究佔 5%[FF1B?] 語言教學與學習佔 40%。如果勉強把翻譯研究歸入語言教學與學習 [FF0C?] 然後把其餘四個領域 [FF08?] 一般語言學 [FF0C?] 心理語言學與神經語言學 [FF0C?] 語料庫語言學與計算語言學 [FF09?] 視為廣義的語言學 [FF0C?] 則全部計畫案約有 55% 是廣義的語言學 [FF0C?] 而語言教學與學習佔 45%。

以上的數字是純就研究領域而分類 [FF0C?] 但究竟這些計畫案是研究哪些語言呢 [FF1F?] 可以想見應該是以國內的國語/華語以及英語、歐洲語言、鄰近的日語韓語為大宗 [FF0C?] 因為我們的外語系或外語學院主要的教學對象就是這些強勢語言。台灣的南島語的研究 [FF0C?] 根據徐嘉慧教授等幾位 ex:key:2017 的著作 [FF0C?] 共有 145 件計畫案 [FF0C?] 也就是在 2006~2015 年間每年約有 15 件計畫。當然根據上面提到的推估法 [FF0C?] 實際上的研究人力應該不只如此。同一時間 [FF0C?] 東南亞語言的研究計畫案有 23 件 [FF0C?] 為數甚少 [FF0C?] 也可以理解。唯東南亞來的新住民早已經有了下一代 [FF0C?] 人口在 2018 年底已經逼近 20 萬人 [FF0C?] 加上近幾年政府推動的新南向政策 [FF0C?] 明白宣示協助新住民利用其語言文化之優勢 [FF0C?] 取得相關證照與就業 [FF08?] 如母語教學、觀光等 [FF09?][FF1B?] 鼓勵大專院校開設南向專業科系或學程 [FF0C?] 給予具南向語言優勢的學生加分錄取 [FF0C?] 培育第二代新住民為南向種子。這個地區豐富的語

言在下一個十年可望逐漸成為台灣語言學界關注的對象 [FF0C?] 擴大語言學更寬廣的研究視野 [FF0C?] 為台灣語言學的發展注入新的活力 [FF01?]

### 0.1.12 摘要與結論

語言的活動是人生際遇的基本配備 [FF0C?] 語言的研究也因而觸及人文社會以及認知 [FF0C?] 資訊或自然科學有關的很多領域。本書涵蓋的許多章節已經多少透露出這樣的訊息 [FF0C?] 值得年輕一代的學子踏上這一趟驚奇之旅。在本章我們引介語言的本質 [FF0C?] 包括下面幾個重要的面向 [FF1A?]

1. 人類的自然語言跟動物的溝通方式本質上的差異 [FF1A?] 遙指能力指語言或動物的溝通是否有能力指向任何一個遙遠的事物 [FF1B?] 反射能力是指語言不但可以談外界的事物 [FF0C?] 也可以談論語言本身。滋生力是指語言有能力表達許許多多幾近無窮的信號。傳統傳承是指人類是從他/她的父母或長者或周遭習得語言 [FF0C?] 每一代都要重新學習。二元結構指語言含有兩層結構 [FF0C?] 一層是沒有意義的個別的而且有限的語音 [FF1B?] 另一層是把這些個別的、有限的、無意義的語音組合成為有意義的單詞 [FF0C?] 有意義的句子。這是語言的數位性格 [FF1B?] 任何動物的溝通系統都沒有這樣的二元結構或數位性格。

2. 語言是一種複雜且具有適應力能「自我組織」的系統 [FF1A?] 自我組織指的是從最初的無序狀態中 [FF0C?] 藉著各部分之間的局部相互作用 [FF0C?] 加上逆向的回饋 [FF0C?] 而產生一個整體變得有序或很有協調作用的過程。這個過程具自發性 [FF0C?] 沒有任何中介主導 [FF0C?] 也沒有任何系統內部或外部的系統在控制。「複雜」是指參與系統的運作的個體非常多 [FF0C?]「調適」是指系統隨時變化 [FF0C?] 以適應新的情境。自我組織的系統是個非線性系統 [FF0C?] 因為它整體的特性具有局部成分所沒有的特性。在一個非線性的動態系統中 [FF0C?] 結構會越來越趨複雜 [FF1B?] 而且結構的形成不假借任何中樞指揮居間策畫 [FF0C?]

而是因為語言使用者彼此有類似的目標 [FF0C?] 有同樣的認知策略 [FF0C?] 而且彼此調適 [FF0C?] 最後形成一個系統。

3. 語言是個動態系統 [FF0C?] 經常變化; 但語言的變化一定是你我對某一個初始條件有某種敏感度。語言的變化往往不可預測 [FF0C?] 因為要能預測變化一定要對導致變化的初始條件有相當程度的了解 [FF0C?] 而我們沒有能力事先了解初始條件。有時候 [FF0C?] 微小的變動 [FF0C?] 最後可能造成明顯的語言變化。因此語言的變化基本上是個混沌現象 [FF0C?] 指的是一個非線性的動態 [FF0C?] 一開始看似毫無章法的現象 [FF0C?] 最後卻變成社群所普遍接受的有規律性的結構。具有混沌現象的系統在發展過程中很容易受外在因素的波動 [FF0C?] 也因此 [FF0C?] 它的演變方向通常無法預測。

4. 語言的時間性 [FF1A?] 指人與人在互動時所講的話語是在時間之流上逐次展開的音聲現象。對話時 [FF0C?] 每個人都有話語權。話語權是協商的結果。也就是在聽話者默許 [FF0C?] 暗示或同意之下 [FF0C?] 講話者才有話語權。由於大腦天生的預測性格 [FF0C?] 聽話者有能力預知何時對方會結束話語 [FF0C?] 輪到自己有話語權 [FF0C?] 而轉而成為講話者。這個預知能力表現在對話或行為的預測方面。聽話者利用他/她整體的語言知識 [FF0C?] 包含話語的聲調、節奏 [FF0C?] 投射對方的話語的走向 [FF0C?] 從而接話 [FF0C?] 或插話 [FF0C?] 或幫對方完成話語。了解語言的時間性與大腦的投射性格才比較容易了解語法系統如何呈現。

5. 吸引子與語言的複雜性 [FF1A?] 大部分的人描述某種情景時通常都趨向使用某一種句型 [FF0C?] 學者稱為吸引子的句型 [FF0C?] 就是相對比較穩定的句型。有了這些穩定的句型才使得語言教學成為可能。但除了穩定的句型 [FF0C?] 也有少數一些人使用其他不同的 [FF0C?] 比較不穩定的講法。這是任何複雜而有調適性格的系統必然的現象 [FF1A?] 有穩定的結構的一面 [FF0C?] 但也有變異之處。變異的結構可能之後反而成為穩定的結構 [FF1B?] 而本來穩定的結構也隨著時間

的變化成為微不足道的變異。

6. 語言 [FF08?] 的使用 [FF09?] 基本上只是個「 素描 」[FF1A?] 由於語言的社會性 [FF0C?] 人類知覺認知系統的不對稱性 [FF0C?] 說話者使用語言表達情境時往往習慣性的只勾勒一個很粗淺的素描 [FF0C?] 而聽話者通常能自動填補話語不足之處 [FF0C?] 就像畫素描畫一樣 [FF0C?] 很自然地會略去許許多多的細節。語言學者感興趣的議題之一是弄清楚到底人類認知系統的不對稱性如何迫使我們很自然地省略某些成分不講 [FF0C?] 但也必定保留某些成分。

## 0.2 第十五章語料庫與語言研究的實證方法

謝舒凱

國立臺灣大學語言學研究所

15.1 背景

15.1.1 相關學門分支

15.1.2 實證方法

15.1.3 語料庫與語言資源

15.2 語料庫工具

15.2.1 整體式網路服務

15.2.2 語料庫程式設計

15.3 語言研究的實證方法

15.3.1 程序

15.3.2 方法論上的考量

15.4 應用

15.5 總結與摘要

15.6 參考文獻

"*Whereof one cannot speak, thereof one must be silent*"

— ***Ludwig Wittgenstein***

## 0.2.1 15.1 背景

語言是個屬於人類的內在複雜系統 [FF08?] 參見第一章 [FF09?]。我們想要探究語言 [FF0C?] 可以立基於內在的直覺語感 [FF0C?] 或是從語言習得的歷程、或語言實際使用與溝通的情境中去觀察與剖析。從語言學的發展歷史來看 [FF0C?] 我們可以發現近二十年來 [FF0C?] 實證與量化研究趨勢 (quantitative/empirical turn) (Luodonpää-Manni, Penttilä, and Viimaranta (2017)) 漸漸的受到重視。這個背景可以說是由多個因素所共同造成 [FF0C?] 包括理論範式的變遷、語料庫的興起與語言科技的迅速發展。

從認知語言學家 (**Langacker1987**) 開始 [FF0C?] 許多學者開始意識到語言結構與意義是由使用中形塑與突現 [FF0C?] 這個想法可以用「基於使用的語言學」(usage-based linguistics) 來稱呼 (Diessel (2017))。這個觀點下 [FF0C?] 語言是個流動的範疇與限制的系統 [FF0C?] 由實際使用中各種高等認知過程交互作用下不斷地重新結構與組織。因此 [FF0C?] 語言學家對於語言使用的工具與方法論就越來越重視。如何掌握語言使用的實際資料 [FF0C?] 也促發了理論與語料的密切結合。本章節要介紹的是語言學研究的實證方法 [FF0C?] 特別是以語料庫為本的取徑。

### 15.1.1 相關學門分支

從語言學史上看 [FF0C?] 使用數學與統計方法探究語言並不是什麼新的想法。談到實證方法 [FF0C?] 我們可以先區分幾個相關且相互交疊的領域 [FF1A?]

- 計量語言學 (quantitative linguistics) 重視的是語言定律 (language law) 的發現。如「齊夫定律」(zipf law) (Zipf 1949) 發現的是在自然語言的語料庫裡單

詞出現的頻率與它在頻率表中的排名成反比 [FF1B?] 曼則拉—阿圖曼定律 (Menzerath-Altmann law) 則發現了語言表達的內部單位長度與其組成的結構單位長度成反比 [FF0C?] 例如句子越長 [FF0C?] 組成它的子句就越短。

- 語料庫語言學 (corpus linguistics) 著重在語言樣本的蒐集、標記與分析。在 1960 年代 [FF0C?] 社會語言學家 William Labov 即開始推動社會語言學的實證研究 [FF1B?] 與此同時 [FF0C?] 第一個英語語料庫也誕生。語料庫可以說是語言使用的「樣本」[FF0C?] 我們可以藉由「樣本」的分析與觀察 [FF0C?] 進而預測與理解語言的本質。也因為「樣本」的概念 [FF0C?] 很自然的與統計方法有密切的連結。語料庫可以說是語言研究實證方法的最大催生力量 [FF0C?] 也是語言學研究社群在近年來成長最快的一個分支。

- 計算語言學 (computational linguistics) 則試圖建立自然語言的計算模型 [FF0C?] 除了可用來模擬人類的語言行為 [FF0C?] 也可以結合語言科技應用 [FF0C?] 讓機器得以學習處理和理解人類的自然語言。

### **15.1.2** 實證方法

以上提到這些學門之間的相同點在於實證方法的重視。實證方法的優點之一是提供較高程度的客觀性、可比較性與可重製性 (reproducible)。可重製性的定義雖然仍有爭議 [FF0C?] 不過大體上說的是一種科學精神 [FF1A?] 在給定相同的樣本數據、假說、實驗設計甚至程式工具 [FF0C?] 我們應該可以預期得到一致的估計參數或結果。之前的語言學研究中因為理論側重不同 [FF0C?] 語料取得與使用的技術也不成熟 [FF0C?] 實證方法顯得比較不被看重 [FF0C?] 甚至有所排斥。但是放在科學研究的進展來看 [FF0C?] 可重製的實驗可以推動理論進步 [FF0C?] 降低研究發現是偶發、隨選或造假的風險 [FF0C?] 因此在語言研究社區中也慢慢為人接受。那麼實證方法與思維要能實際作用 [FF0C?] 其中有個重要的關鍵在於「資

料」或稱「數據」(data)。和語言產出、使用、分析有關的資料我們可以稱之為語言資料 (linguistic data)[FF0C?] 其中最為廣泛使用與分析的是語言自然使用的語料 (corpus data)[FF0C?] 和用語言田野工作或相關實驗導出的資料 (elicited data)。隨著當前的語言科技進展倍速 [FF0C?] 在語料的錄製、辨識、前置處理、標記、並列、校準、分析建模、視覺化等方面 [FF0C?] 實證方法在語料處理與分析上的應用可以說是已經達到相當的成就。

### **15.1.3** 語料庫與語言資源

任何以語言溝通系統為主的數位化資源 [FF0C?] 我們都可以稱之為「語言資源」(language resource)。其中作為語言使用的代表性樣本的語料庫可以說是最為典型的一種語言資源。語料庫 (corpus) 與語料庫為本 (corpus-based) 的研究 [FF0C?] 也可以說是當前語言學研究的核心的方法之一。它也是語言實證研究的主要的工具之一。那麼什麼是語料庫呢 [FF1F?] 語料庫語言學家 (**Gries2019**) 認為一個典型的語料庫需符合幾個條件 [FF1A?](1). (口語或文本) 語料符合機讀格式 (machine-readable)[FF0C?](2). 語料是在自然溝通語境下蒐集得到的 [FF0C?](3). 就蒐集對象來說要考量到樣本代表性 [FF0C?] 在文類、文體變異也要取得平衡。(4). 要能夠被編制成便利語言學分析的樣式。

當前大部分語言相關的研究在不同程度範圍內 [FF0C?] 都使用到語料庫資源與工具。此外 [FF0C?] 從時代背景來看 [FF0C?] 隨著社交媒體與社會網路的發展 [FF0C?] 非結構性的文本資料所占比例已遠超過結構性的表格性資料 [FF0C?] 使得文本的語言分析在資訊發展中的角色也顯得愈來愈吃重 [FF0C?] 連帶的語料庫與分析方法更擴及到語言與教學研究、社會科學、神經心理與認知科學研究上。

近年來光在語料庫語言學的發展進程上 [FF0C?] 就產生了許多新的趨勢。例如在來源取得方面 [FF0C?] 傳統手製語料庫的方式 [FF0C?] 已經轉向半自動建立之

巨型語料庫 [FF0C?] 一直到網路作為語料庫 (web as corpus) 之實現。特別是社會媒體網路可能是一個最為相關且有趣的語料來源。網路時代的線上生活 [FF0C?] 不僅提供活生生的當代語言使用資料 [FF0C?] 更重要的是還有彼此相互連結的人際溝通意圖訊息。在以前因礙於語料取得困難 [FF0C?] 傳統的語料庫概念比較假定文本的封閉靜態 [FF0C?] 使得語言行為呈現單調 [FF0C?] 樣本亦常常受限代表性的問題而欠缺說服力。諸如多模態語料 (multimodality)、領域語言 (sublanguage)、個人習語 (idiolect)、領域知識本體 (domain ontology) 等研究與應用 [FF0C?] 都難以展開至一定規模 [FF0C?] 而今語料的大量可得性像是開啟了潘朵拉的盒子。這些巨量與多元的語料開始讓我們有機會重新思考一些語言學的幾個基本爭論點 [FF1A?] 語料是什麼 [FF1F?] 語料與語言理論的關係是什麼 [FF1F?] 語言學習者需要的語料是什麼 [FF1F?] 有哪些 [FF1F?] 等等。

此外 [FF0C?] 在語料的多元方面 [FF0C?] 語言資源除了以語言使用「共面」(syntagmatic) 為主的語料庫之外 [FF0C?] 尚有以「立面」(paradigmatic) 詞彙資源為主者 [FF0C?] 像是同義詞辭林 (Thesaurus)、詞彙網路 (WordNet)、框架語意 (FrameNet) 等詞彙與詞典資源。目前在正體字華語社群中 [FF0C?] 以中文詞彙網路 (Chinese Wordnet) (Huang et al. 2010) 和繁體知網(eHowNet) 為分析與標記最完整的詞彙知識庫。以中文詞彙網路為例 [FF0C?] 它的設計理念是在完整的詞彙知識系統下 [FF0C?] 兼顧詞義與詞義關係的精確表達與語言科技應用。這項資源裡有兩項重要的元素 [FF1A?] 一是以詞義為據的詞彙分組 (即所謂的同義詞集 (synset))[FF0C?] 另一個就是聯繫同義詞集的語意關係 (semantic relations)。以同義詞集為節點 [FF0C?] 透過語意關係相互聯繫 [FF0C?] 就形成了表徵詞彙意義及其關係的語意網路 [FF0C?] 也可視為是以詞義與語意關係為經緯建立的人類語言知識表達基本架構模擬。建構完成的詞匯語意網 [FF0C?] 一方面可以作為語言學研究的素材 [FF0C?] 另一方面在各項資訊應用上都可當成是重要元件。

### 0.2.2 15.2 語料庫工具

如前面提到的 [FF0C?] 語料庫是最為普遍使用的一種語言資源。不過我們要瞭解到 [FF0C?] 語料庫本身並不會直接提供給我們資訊。我們需要有假說 [FF0C?] 有適合的工具。而實證與經驗方法的使用 [FF0C?] 自然地同步會發展相應的工具集 (toolkits)。我們可以就不同目的把它們分成不同的類型。從系統實作的角度來說 [FF0C?]Kosem and Kosem (2011) 把語料庫工具類型分成幾種對比 [FF1A?] 單機 vs. 線上工具、與特定語料庫有關 vs. 獨立於任何語料庫、簡易 vs. 進階工具等等。就內容來說 [FF0C?] 語料庫工具類型包括語料庫的建構、預處理、標記與分析等面向 [FF0C?] 如下所述:

- 建構

所謂的建構包括了語料的搜集 (collection)、清理 (cleaning)、編製索引 (index) 與儲存 (storage) 工作。一開始的語料搜集工作涉及了如何透過程式工具取得語料 [FF0C?] 以及所搜集到的語料樣本在各式文類 (genre) 與文體 (mode) 的平衡性考量 [FF1B?] 清理工作則取決於不同研究目的 [FF0C?] 對於所搜集到之樣本進行訊息保留或清理。編製索引的目的是期許在日後讓語料庫的應用程式可以快速檢索與搜尋 [FF0C?] 市面上有不少的可能方案如開源的搜尋引擎 ElasticSearch、Apache Lucene[FF0C?] 語料庫語言學社群中也開發了如 Emdros 和 Corpus WorkBench 等平台 [FF0C?] 後者是目前比較受到歡迎與使用的工具平台。

- 處理

包括切符 (tokenization)、【中文】斷詞 (或分詞) (word segmentation)、詞類自動標記 (POS tagging)、句法剖析 (parsing)、其他語意或語用等語言學訊息的人工標記 (annotation platform) 或自動標記工具 (tagger) 等。此部分常與計算語言學的研究工作重疊。

• 標記

標記任務指的是以人工或半自動機器輔助的方式 [FF0C?] 將不同層次與面向的語言訊息標在語言單位與序列上。標記好的語料對於語言分析與機器學習有很大的助益。較著名的協作標記工具 (teamware) 有 GATE、MAE、BRAT、WebAnno(Inception)、ANNIS 等等。除了以上發展較久的標記系統 [FF0C?] 晚近更有將標記整合到程式處理的架構如 Prodigy + spacy。

• 分析

這個部分可以說是語料庫方法的核心成分 [FF0C?] 包括對語料進行瀏覽、統計、模式抽取等工作 [FF0C?] 依照不同的研究目的而有區別。廣義的說 [FF0C?] 我們可以綜整成幾個方向 [FF1A?]

(14)

. 頻率 [FF1A?] 所要探究的語言單位 [FF08?] 字、詞、結構等 [FF09?] 的類型 (type) 與實例 (token) 的發生次數計量。

(15)

. 分佈 (distribution) 與聚散 (dispersion)[FF1A?] 探究語言單位類型 (type) 與實例 (token) 在語言樣本中的分佈與聚散程度。

(16)

. 共現搭配 (coll(oc|ig)a|ostruc)tion))[FF1A?] 探究語言單位或結構之間在類似語境下共同搭配出現的現象。

(17)

. 關鍵顯著 (keyness)[FF1A?] 探究語言單位在語料庫中的顯著代表性。如關鍵字等。

(18)

. 視覺化呈現 (visualization) 與圖形特徵計算 [FF1A?] 探究語言單位之間的全域連結 [FF0C?] 常結合統計計算與計算語言學模型 [FF0C?] 如叢聚分析 (cluster analysis)、語意相似度計算 (semantic similarity) 、網路分析 (network analysis) 等等。

分析輔助工具上 [FF0C?] 單機版的語料分析工具是歷史最悠久的工具 [FF0C?] 例如免費的 Ant*系列[FF0C?] 到目前為止都是許多語料庫研究者目前仍在使用的分析工具。特別是 Ant* 系列除了提供常用的文本分析工具 [FF0C?] 還提供中日文自動斷詞與詞類標記、語步 (move) 標記分析、字元轉換、字彙難易度、語言變異等工具 [FF0C?] 對於入門的讀者相當實用 [FF0C?] 值得推薦。此外 [FF0C?] 英國蘭卡斯特大學語料庫團隊近年也推出了自由軟體 #LancsBox[FF0C?] 對於搭配詞的網路提供了統計公式調整與視覺化的呈現 [FF0C?] 對於語料量不大的研究者來說 [FF0C?] 是一個不錯的分析工具。

### **15.2.1** 整體式網路服務

上述的單機版工具雖然小巧方便 [FF0C?] 但是隨著語料的可得性指數增長 [FF0C?] 在個人電腦上處理大數據常常會因為計算資源的捉襟見肘而難以順利進行。如果要處理的語料量較大 [FF0C?] 就需要使用線上的語料分析系統。這些系統提供蒐集與處理好的語料 [FF0C?] 並提供工具讓人可以線上使用 [FF0C?] 不需耗費個人電腦的資源。其中比較著名的有當代美語語料庫 COCA、和 Word Sketch Engine。此外 [FF0C?] 大部分的讀者比較不熟悉的是歐洲的傳統 [FF0C?] 比方說瑞典哥登堡大學開發的 Språkbanken (the Swedish Language Bank 甚至整合了如 FrameNet 等詞彙語意資源與標記 [FF0C?] 使整合性語料庫語言學更向前邁進。而在臺灣正體中文的語料庫研究社群中 [FF0C?] 向來以使用中研院平衡語料庫

ASBC 為主。不過 [FF0C?] 隨著語料庫停止收錄更新 [FF0C?] 比較難以符應語言使用的當代特性 [FF0C?] 加上語料標記的多元化也漸漸成為整合研究的需要 [FF0C?] 已有不少規模不大但具有不同特色的漢語語料庫公開提供外界使用 [FF0C?] 如臺灣師範大學的華語為第二語口語語料庫、政治大學漢語口語語料庫、臺灣大學語言所 LOPE 實驗室開發的開放語料與搜尋系統 COPENS 和動態更新的批踢踢語料庫等等。

這些系統的開發與開放 [FF0C?] 也促進了語料庫語言學分析工具的創新 [FF0C?] 讓語言的實證研究更為有趣。比方說 Word Sketch Engine 利用語法規則可以呈現搜尋語詞的語法搭配詞。如下圖以「吃」為例 [FF0C?] 我們可以觀察到「吃」的主詞、受詞、修飾語等分佈 [FF1B?] #LancsBox 則提供視覺化工具來呈現搭配詞與搭配詞類 (colligation) 的網路 [FF1B?] 批踢踢語料庫則致力於建構對話與功能語料庫的架構設計。

圖 15-1

圖 15-2

**15.2.2** 語料庫程式設計

不過 [FF0C?] 隨著資料的劇增與資訊科技的快速發展 [FF0C?] 在許多時候現有的語料庫系統與工具已無法符應研究需求。比方說目前在單機版執行的語料庫工具處理大型資料的能力有限 [FF0C?] 語料搜尋與處理的趨勢還是以雲端運作為主。而隨著語言分析的層次與廣度 [FF0C?] 目前語料庫系統提供的功能常常不敷使用 [FF0C?] 使用者或研究者即便有創新的觀點 [FF0C?] 也很難期待現有的語料工具開發者能夠迅速地同步更新。此外 [FF0C?] 目前的數位資料累積與增長的速度已遠遠倍增於人類史上的任何階段 [FF0C?] 這樣一種巨量資料風潮也改變了人文社會與自然科學研究的面貌在此背景下 [FF0C?] 直接學習「語料

庫程式設計」(corpus programming)[FF0C?] 自行開發與解決手邊的研究與實務問題 [FF0C?] 同時也可回饋社群 [FF0C?] 造就更廣泛的知識進化與傳播便成了新的學習與研究趨勢。所謂以程式方法研究語料庫 [FF0C?] 簡單來說就是碰到問題與需要時可以用編寫程式的方法解決。隨著學習工具與網路集體學習的快速發展 [FF0C?] 目前學習編寫程式的支援環境與資源已相當友善、普及、甚至目不暇給。有興趣的讀者可以參考我們在 2018 年舉辦的語料庫程式設計入門工作坊(https://github.com/lopentu/BestPracticeInCorpusProgramming)。以下我們以 R 語言為例 [FF0C?] 使用 Levshina (2015) 提供的數據來當成簡單的展演例子。

```
require(Rling)
data(ldt)
# 100 selected words for the Lexical Decision Task
```

首先我們載入一個從 English Lexicon Project Balota et al. (2007) 的實驗語料 [FF0C?] 包含了隨機選出的 100 個英文詞彙 [FF0C?] 以及它們的詞長、平均反應時間和在語料庫中的使用頻率。之後很快可以以程式 summary(ldt$Mean_RT) 得到基本描述統計 [FF08?] 以平均反應時間為例 [FF09?]: 564.18, 713.1125, 784.94, 808.2533, 905.2, 1458.75[FF0C?] 也可以迅速作圖觀察變量之間的關聯與分佈。

圖 15-3

或跑統計檢驗 shapiro.test(ldt$Mean_RT)[FF1B?]

```
##
## Shapiro-Wilk normality test
##
## data: ldt$Mean_RT
## W = 0.92, p-value = 0.00001
```

或算出相關係數 (cor(ldt$Mean_RT, ldt$Length) = 0.6147) 得知字長與反應時間的

正相關 [FF1B?] 或跑散佈圖與跑迴歸分析等等。

圖 15-4

### 0.2.3 15.3 語言研究的實證方法

當然 [FF0C?] 會寫程式不等於能夠做研究。接下來我們就具體一點的來談如何運用實證方法 (empirical method) 來探究語言。不過我得預設讀者具備基礎的統計知識 [FF0C?] 這也是實證方法的基本先備知識之一。因為我們採取是經驗性的觀點來分析語言 [FF0C?] 從抽象的理論到具體的觀察之間是有逐層的處理步驟可以參照。(Manheim et al. 2008) 用以下的圖示說明了在經驗實證方法中概念 (concept)、變項 (variable) 與指標量度 (indicator/values) 之間的關係。

圖 15-5

左右欄是一個理論構想與實證方法的對應。首先我們對於關心的語言現象有理論預想 [FF0C?] 對應到右邊就是指某概念與某概念之間有某種關連。理論會有假說 (hypothesis)[FF0C?] 但是在實證研究中必須經過「操作化」**(operationalized)** 的轉換過程。所謂可操作化是一個選擇可觀察到的具體現象來表徵抽象概念的過程。為了做到這件事 [FF0C?] 我們必須找到量度工具用來量化地表達我們假說中的概念。再者我們要了解 [FF0C?] 經驗研究中所謂的「可觀察到的」[FF0C?] 指的是要分析的現象特徵是可以藉由量度工具來賦值的。這樣說有點抽象 [FF0C?] 舉例來說 [FF0C?] 我們有興趣探究到底文法系統是內在的模組化系統 [FF0C?] 還是外在學習突現出來動態系統的這麼大的主題。(Goodman 1997) 提出了可以觀察「詞彙量」(vocabulary size) 和 16-30 個月大的小孩的「文法發展」(grammatical development) 這兩個變項之間的關係 [FF0C?] 並假設了它們之間存在著正相關。他們將變項作的「操作化」就是詞彙量以「受試小孩子所產出的詞彙數量」來量度 [FF0C?] 而文法發展則是以「所學到的構式數量」(作者先預先定義了 37 種構

式)。假說的右邊對應就是將假說操作化為變量之間的關聯。接著我們會有所謂作業假說 (working hypothesis)[FF0C?] 這與以下會提到的「探索性分析」密切關聯。可以把它當成是一種臨時接受的假說 [FF0C?] 希望透過探索分析得到驗證或拒絕 [FF0C?] 作為進一步研究的基礎。最後就是實際的觀察值的蒐集代入。到這裡我們可以看出 [FF0C?] 經驗方法的好處之一在於如果實驗資料是公開或是可模擬的 [FF0C?] 我們就可以重製 (reproduce) 之前的實驗 [FF0C?] 並在其基礎下往前繼續深入探究。

### **15.3.1** 程序

在實際的資料分析流程上 [FF0C?] 有一個一步一步來的慣例規約 (protocol)。下圖大概表示了這樣的流程。但要注意每個模組裡都有更細緻的子模組 [FF0C?] 模組之間的運作也並不是單向式的 [FF0C?] 而是「來來回回」(back and forth) 的互動。每個步驟都有相應需要注意的地方。

圖 15-6: 語料實證分析流程

假設我們對於中文單音節詞 (就是由單一漢字所構成的詞所組成的詞) 的心理處理歷程有興趣 [FF0C?] 並假設不同的使用頻率與反應時間會有造成不同的效應。那麼採取經驗方法的話會涉及哪些步驟呢 [FF1F?] 以下我們用 Sun et al. (2018) 這篇論文所提供的現代漢語心理詞庫資料 [FF0C?] 來重製部分實驗 (見以下引自原文的表格) 與說明實證經驗方法的程序與好處。

(**SunEtAl2018**) 這篇文章介紹了一個開放下載的中文心理詞庫 (Chinese Lexicon Database, CLD. http://www.chineselexicaldatabase.com)。這個詞庫提供了許多現代漢語詞彙的各種心理實驗數據 [FF0C?] 方便研究者進行不同的語言實證研究。不過這個資源是基於簡體中文 [FF0C?] 假定我們想先重製詞彙判斷作業反應時間 (lexical decision latencies) 與從語料庫抽取出來的頻率之間的關聯實驗 [FF0C?] 並

看看在正體中文是否會得到類似結果 [FF0C?] 或觀察不同時間取得的頻率 [FF0C?] 是否結果相似。

資料取得

我們首先從 CLD 取出資料 [FF0C?] 這個 CLD 詞庫 (v.2.1) 中包含了 48,644 個漢語詞彙 [FF0C?] 及其在 269 個變項的數值。利用我們開發的 R 套件 lexicoR (https://lopentu.github.io/lexicoR/articles/databases.html[FF09?][FF0C?] 很方便的即可取得所有的資料。以變數項目來說 [FF0C?] 就有字形筆畫數、使用頻率、字長、音節、熟悉度、聲調、鄰近密度量度、詞彙判斷作業反應時間 (RT)、字詞命名反應時間、資訊熵值量度等等。

前處理

因為我們要重製部分實驗 [FF0C?] 因此不是所有的變量都用得到。在此我們只取出單音節詞、與作為回應變量 (response variable) 的反應時間 ( RT, 毫秒單位) 、作為解釋變量 (explanatory variable) 的其他幾個變量 [FF08?] 如轉成正體字的筆畫數、從不同時期批踢踢語料庫計算出來的詞頻等等 [FF09?]。此外 [FF0C?] 我們有時候也需要改變量名稱增加可讀性。經過前處理之後的資料大概長這樣 [FF1A?]

Ch 4_2009

—58

—

—

- 啊

哎56

哀57

唉68

埃72

挨68

挨68

癌59

矮60

艾59

爱49

碍67

安51

鞍65

谙82

注意的部分 [FF0C?] 通常可以藉由視覺化技術更好的呈現與挖掘。比方說 [FF0C?] 我們需要先看看 CLD 中的在詞彙反應作業任務中 [FF0C?] 頻率、平均反應時間與筆畫數之間共線性的檢查 [FF0C?] 可以用 multi-panel 散佈圖和皮爾森關聯係數。

統計建模

我們大概知道了變量的基本量化特徵 [FF0C?] 變項之間的關聯 [FF0C?] 如果我們要進一步了解資料底層的模式與互動機制 [FF0C?] 我們就需要使用統計模型。所謂統計「建模」(modeling) 其實就是把我們的對於現象的觀察心得轉譯成數學式子。以反應時間與使用頻率兩個變量來說 [FF0C?] 我們可以建立線性迴歸模型 [FF0C?] 得到

```
##
## Call:
## lm(formula = Mean_RT ~ LogFreq, data = LDT1)
##
## Residuals:
## Min 1Q Median 3Q Max
## -145 -38 -7 31 291
##
## Coefficients:
## Estimate Std. Error t value
## (Intercept) 626.37 1.81 345
## LogFreq -15.73 0.56 -28
## Pr(>|t|)
## (Intercept) <0.0000000000000002
## LogFreq <0.0000000000000002
```

##

## Residual standard error: 56 on 2414 degrees of freedom

## Multiple R-squared: 0.24, Adjusted R-squared: 0.24

## F-statistic: 7.8e+02 on 1 and 2414 DF, p-value: <0.0000000000000002

因為語言現象總是涉及了許多變項。我們如果想要看變項之間的關係 [FF0C?] 與其對於現象/模型的影響 [FF0C?] 通常引入複迴歸分析 (multiple regression)[FF0C?] 這裏就不再細談。

### **15.3.2** 方法論上的考量

在第 1 節中我們提到的基於使用的語言觀點中 [FF0C?] 頻率是個極為關鍵的概念。Gries and Ellis (2015) 特別指出語言使用的頻率計算上 [FF0C?] 有許多值得注意的地方 [FF1A?]

- 操作化不免會簡化問題的本質。
- 好語料決定大部分模型的結果。
- 分析的透明性與重製實驗的設計。

儘管採取實證方法 [FF0C?] 許多研究僅僅提供統計檢定與最後實驗成果 [FF0C?] 不提供實驗數據而無法重製實驗 [FF0C?] 這也是造成許多研究要累積往前可能的障礙之一。此外 [FF0C?] 實證與量化方法可以說是取得很大的成就。但是我們同時需要注意單一思維與研究取徑總是有其相當的限制 [FF0C?] 比方說 [FF1A?]

- 統計上的關聯 (correlation) 並不保證因果 (causality)。
- 避免「擇優挑選」(cherry-picking) 的建模過程。

- 數據的取得與種種限制 [FF0C?] 使得語料庫與認知的對應 (from-corpus-to-cognition) 關係建立要相當謹慎。
- 中文語料的斷詞問題 [FF0C?] 在巨量語料庫時代因為失去人工檢查的可能 [FF0C?] 隨著語料的擴增 [FF0C?] 雖然微量的斷詞錯誤累積後仍會導致分析的失準。

### 0.2.4 15.4 應用

有了語料庫 [FF0C?] 我們觀察語言多了更具備經驗、實證性的角度。此外 [FF0C?] 也容易與資料科學、自然語言處理 [FF08?] 計算語言學 [FF09?] 等資料密集的學門有進一步的連結與互動。以下我們將以語言的變遷與計算語言方法來說明。

我們之前都是利用語料庫來研究語言的「橫的軸向」[FF0C?] 我們也可以用不同時期的語料來做語言的「縱的軸向」探究 [FF0C?] 以實證語料為本去看語言在不同層次的變遷、演化等問題。隨著巨量歷時語料的增加 [FF0C?] 許多學者例如 Michel et al. (2011) 開始嘗試用 Google 從 1800 年開始數位化的書籍資料進行語詞的量化觀察 [FF0C?] 並探究語詞的歷時使用分佈與「文法演化」的問題。這樣的資料導向分析法 [FF0C?] 更便利了跨學門的整合觀察 [FF0C?] 我們有機會從語言、文化與社會的角度來了解不同的現象 [FF08?] 所謂的「文化資訊學」(culturomics) 也因此誕生 [FF09?]。

我們可以「老婆、太太、愛人」為例 [FF0C?] 搜尋並取得 Google Book ngram 跨兩百多年的書籍 N 連詞 (n-gram) 使用頻率資料 [FF0C?] 如下圖顯示 [FF0C?] 可以看到太太的稱謂在 1940-50 年左右達到高峰 [FF0C?] 對照國民政府時期的官太太文化 [FF0C?] 也許可以得到相應的理解。

圖 15-7

不過正如 Brezina (2018) 的提醒 [FF0C?] 如果沒有進入脈絡去觀察 (如我們先前提到的 concordance 環境)[FF0C?] 很有可能會有嚴重的誤讀。

晚近在計算語言學的分佈式與分散式語意表徵 [FF08?] 如詞嵌入 word embeddings[FF0C?] 請參見第 16 章 [FF09?] 的進展 [FF0C?] 可以將共現脈絡訊息某個程度地抽象化投射到低維度高稠密的向量空間 [FF0C?] 使得語言的使用可以用向量的表徵與運算來運作。這種基於語境數據的向量表徵 [FF0C?] 對於語言研究也帶來新一波的視野。我們 (Chen and Hsieh, 2019) 利用了中文文史哲歷史文獻 CTEX[FF0C?] 與其他歷時語料建立了漢語跨千年的歷時語料庫 [FF0C?] 方便對於語詞的概念變遷做實證的分析與觀察。下圖的例子 [FF0C?] 是利用這樣的語料庫 [FF0C?] 我們可以利用搭配詞看到「家」的規則多義 (regular polysemy) 隨著時代在「人的關係組合」、「居住處」、「成員」之間的語意變遷。例如從圖 15-9 的家_9 的變化來看 [FF0C?] 這個詞義比較像是「每一＜家＞的拳法皆不相同 [FF0C?] 走步也不相同 [FF0C?] 好像基本的馬步是一定有的。」中的意思 [FF0C?] 而隨著時代演變 [FF0C?] 這樣的用法日漸罕見。

圖 15-8

圖 15-9

## 0.2.5 15.5 總結與摘要

本章介紹了我們要探究語言現象時 [FF0C?] 可以採取的實證研究方法。

語料庫就像是語言使用的收集 [FF0C?] 基於語料庫來從事語言研究 [FF0C?] 不論在語言結構本身 [FF08?] 語音、構詞、句法、語意等 [FF09?][FF0C?] 或是語言的形塑因素 [FF08?] 認知、文化社會、歷史 [FF09?] 等方面都有了長足的進展。此外 [FF0C?] 語料庫對於資料科學 (data science)、文本採礦 (text mining) 與自然語言處理 (natural language processing) 等語言相關的科技發展有著密切的關係。

讀者若對於語料庫程式設計有興趣 [FF0C?] 以下是可以參考的書籍。

- (Levshina 2015) 的 **How to do linguistics with R** 是不錯的入門書。
- (Brezina 2018) 和 (Gries 2009, 2010, 2016) 則是語料庫統計的一般性介紹。可以和 #LancsBox 的線上課程一起搭配學習 [FF08?] http://corpora.lancs.ac.uk/lancsbox/materials.php[FF09?]。
- (Winter 2019) **Statistics for Linguists: An Introduction Using R** 則是語料庫統計學的入門好書。