理論語言學研究於大型語言模型熱潮中的定位

陳宗穎*

一、前言

語言學研究在經過 1950-60 年發生的「認知革命」後，主要任務之一是以科學的方式提出理論、測試假設、並企圖解釋人類大腦的語言知識與語言處理機制。隨著近三十年現代科技的發展，理論語言學家得以使用更多的科學工具（如：電腦化的行為實驗、腦波測量、統計／電腦模擬等）探究人類語言知識的本質，也讓語言學成為融合社會科學與自然科學的一門領域。這數十年來，語言學科學研究已經累積了對人類語言一些關鍵的理解：（一）人類語言雖然表面上都是以「線性排列」的方式呈現，這些線性排列卻有著「非線性」的結構。舉例來說：在人類語言裡（包含中文），反身代名詞「（他）自己」指涉的對象並非取決於名詞在同一個句子中的先後順序。（二）人類語言知識包含「線性的排列組合」與「非線性的抽象結構關係」，而後者又以「符號規則」（symbolic rule）的方式運作（詳見本文第三節）。（三）人類在嬰幼兒時期，基於有限的語言輸入就已經開始發展（二）提到的語言知識，而這些語言知識（特別是符號規則）又成為了我們能夠產生與理解人類語言無限可能性的基礎。然而，近年大型語言模型（Large Language Model, LLM）在「模仿」人類語言輸出的技術突破似乎成為挑戰前述語言學研究發現的核心論點。更有學者主張 LLM 代表的就是人類語言理論，而研究 LLM 的語言表現便能開拓我們對人類語言本質的了解（如：Piantadosi, 2024）。這些隨著 LLM 發展而產生的論述可能進一步地邊緣化語言學研究，也是近期語言學領域辯論的重心之一。如〈義大利語言學期刊〉（*Italian Journal of Linguistics*）在今年即出刊了一卷包含理論與計算語言學家觀點的特輯。[1] 筆者期望能在篇幅有限的情況下，強調語言學者加入這些對話的重要性，

* 國立清華大學外國語文學系副教授

[1] 所有文章皆以開放取用形式刊載於：https://www.italian-journal-linguistics.com/2025-2/。

同時也從個人觀點說明目前主流的 LLM「並非」理解人類語言知識架構主要管道的理由以及語言學領域未來可能的發展趨勢。[2]

二、現代主流 LLM 的語言學基礎

目前主流 LLM 的演算法會根據不同的任務導向最佳化，但底層的基礎框架仍然是利用大型的數位資料庫進行權重訓練的人工類神經網路（Artificial Neural Network, ANN）。這些通常具備數千億個參數的 ANN 在經過深度學習函式運算調整參數的權重後，即可根據輸入的提示以及過往的互動內容合理「預測」適當的輸出，進行文字接龍。而 ANN 作為 LLM 的基礎訓練架構，根據上一組詞元標記（token）預測下一個線性輸出的 token 也確實與人類的表層語言現象有所重疊，因為人類語言在表層也是線性排列的。在一個像是「我喜歡你」的簡單句子中，我們可以依照字詞順序定義在動詞「喜歡」之前的名詞就是主詞、之後的名詞就是受詞，而無需考慮在這個句子裡還有任何內部的非線性結構關係。同時，自然語言中，長度簡短且結構簡單語句的使用頻率較高。因此，語言模型以線性的方式處理、預測、並產生自然語言並不一定會與人類日常生活使用語言產生顯著的表層差異。人類的語言理解與輸出也同樣包含預測的成分。舉例來說，當我們聽到「生日」一詞，便會預測接下來出現（或需要輸出）的詞可能會是「快樂」、「蛋糕」、「願望」（亦稱做誘發效應，或 priming effect），而不會是「電燈」、「公車」、「椅子」等詞。心理語言學研究長久下來累積的實驗證據說明人類語言的聲音、詞彙、乃至於語句皆在大腦記憶空間中透過一個大型的網路連結，也支持以網路模擬部分人類語言處理歷程的基礎。在電腦算力以及記憶容量尚無法支持建構大型語言模型的年代，功能學派的語言學家其實就已經提出理論、假設人類語言知識的基礎來自於記憶中海量的語言輸入（如 Bybee, 2001），而抽象的語言知識則是從累積海量語言輸入中自然的「湧現」（亦即是 emergence view）。假設某個語言的母音有長短的差異，而長母音的長度分布大約是 250-350 毫秒、短母音的長度則是分布在 150-250 毫秒。雖然兩種母音的長度分布範圍有所重疊，但大多數的長母音與短母音的長度可能分別在 300 毫秒與 200 毫秒上下，所以在海量的語言輸入中，長短母音兩種抽象的聲音種類就透過不同的統計分布自然地「湧現」了。這樣的理論架構也與 LLM 的訓練有著

[2] 本文只聚焦在人類與 LLM 的語言表現，所以 LLM 本身的應用價值（如：工作效率的提升、圖像辨識與生成、藝術創作等）並不在本文討論的範圍之中。

類似的邏輯：透過海量的資料調整 ANN 之中的節點權重，而最終訓練完成的節點權重代表的就是某種語言統計規律性自然「湧現」的結果。這樣的「巧合」也是大型語言模型獲得一部分語言學家支持的原因之一。

三、特化的抽象符號語言知識與大型語言模型的基礎差異

由於前述的語言理論與語言處理歷程的實驗證據與大型語言模型的基礎架構相同，而 LLM 又在資料量以及算力提升的情況下在語言任務中的表現突飛猛進，有學者（如：Piantadosi, 2024）甚至提倡 LLM 本身即代表了人類語言的理論，而研究 LLM 中自然「湧現」的語言規律性就可以理解人類的語言知識與本質（見 Kodner 等人，2023；Moro，2023 的反對意見），並從根本上邊緣化語言學研究。

然而，語言知識雖然包含了語言記憶中個別語言標記相互連結的大型網路與量化的統計分布，卻不代表語言知識中僅僅包含了這些連結與統計資訊。其實許多功能學派的語言學家也提倡所謂的混合式語言模型（hybrid model），也就是除了大量的語言記憶之外也包含抽象層級的語言知識（如：Pierrehumbert, 2006），而在工程技術方面則可以透過神經符號模型（neurosymbolic model）實踐混合語言模型。抽象知識的特性之一就是可以舉一反三，外推（extrapolate）到訓練資料之外的例子上。一個數學方程式像是 $x + x = 2x$ 即是一種抽象的代數符號知識；無論 x 的數字為何、也無論我們是否曾經看過這個數字，我們知道兩個相同的數字加總就是該數字的倍數。可以外推的代數符號知識是人類有效率地習得語言與進行語言溝通的關鍵。常用的語言學概論教科書都會提到人類語言有著無窮盡的語句可能性，而即便人類語言聲音系統中的類別有限，也沒有任何一個語言會使用到所有的聲音類別與聲音類別組合。若沒有一種可以外推的抽象代數符號知識，我們唯一能夠進行語言溝通的方式就只能像 LLM 將語言輸入與數位資料庫中的海量資料進行比對、或是將數位資料庫中的海量資料進行重新排列組合產生新的輸出。

過往對於抽象語言知識的討論大多聚焦在語句結構層面，但認知科學家 Iris Berent 教授根據她的希伯來語（Hebrew）實驗發現提出人類的音韻心智（phonological mind）也是一種抽象「代數知識」（Berent, 2013）。希伯來語詞彙中有禁止兩個相同子音的限制，但這個限制在有三個子音的詞彙中只針對詞彙中最左邊的兩個子音（如：*sisim；「*」代表「不能出現」或「不合語法」），所以該

限制稱做為「*AAB」[3]。在該實驗中，Berent 發現希伯來語的母語者不但將「*AAB」的規則外推至希伯來語中不存在的詞彙，甚至將該規則外推到不存在於希伯來語中的子音。這也是人類的語言系統與「單純建立於海量資料和 ANN」的 LLM 最廣為人知的差異之一；只有後者會受到訓練資料的限制。在基本架構不變的情況下，主流 LLM 在語言任務中遭遇瓶頸時，最主要的突破口就是進行訓練資料庫的擴展（亦稱為 scaling），但在人類語言可以產生無窮盡語句的可能性之下，全然倚賴擴展並無法以涵蓋所有的表層語言現象的方式納入或呈現人類的語言知識。

除此之外，人類的抽象語言知識似乎是一種經過特化過的能力，也讓人類在學習語言上有著學習特定語言結構先天的「偏好」。在 Culbertson 等人（2012, 2020）知名的「人工語法習得實驗」中（artificial grammar learning），受試者接收某個未知人工語言的語言輸入，而這些輸入隱藏著特定的名詞片語語序規則，像是「名詞—形容詞」、「量詞—名詞」等。在受試者未被告知有這些特定的語言規律性的情況下，實驗發現這些受試者比較容易學習到名詞在前的語序，如「名詞—形容詞」與「名詞—量詞」。這可能是因為名詞帶有名詞片語的核心資訊，所以先處理核心資訊再處理次要資訊（如：名詞的描述與數量）比較符合人類的語言處理歷程。人類在學習抽象語言知識有著先天偏好似乎並不令人意外，因為這代表人當人類不用考慮聲音字詞的所有可能排列組合時，反而可以更快、更有效率地學習到目標語言知識或是處理語言資訊。有趣的是，雖然先天的學習本能在語言學領域常常是辯論的重心（詳見 Berent, 2025），在生物界其實相當常見。Gallistel 等人（1991）主張無論是人類或其他生物都有著特化的認知功能結構可以引導不同群體學習在自然環境中特定的生存法則。作者群引用的其中一個實驗是讓鴿子為了獲得食物或是避免被電擊學習啄咬鑰匙的動作。雖然鴿子很快就學習到啄咬鑰匙與獲得食物的連結，但是卻無法將同樣的動作與避免電擊進行連結。不過若是將避免電擊做為拍打翅膀的「獎勵」，那麼鴿子也同樣很快就可以學習到兩者的連結。這種學習偏好來自於鳥類的生物本能；當鳥類受到威脅時（如：對腳部的電擊），它們會拍翅飛行或至少嘗試飛行，而不是啄咬任何的物品。Gallistel 等人也提到，在自然的生存情境中，我們很難想像任何生物需要在被掠食者攻擊數次才能透過經驗法則從這些「學習輸入」獲得「不逃就會有生命危險」的知識。雖然人類語言學習相較於躲避掠食者的攻擊來說似乎很難相提並論，但若是語言溝通是一種複雜且包含生存知識的訊息傳遞

[3] 當讀者在這裡第一次看到「*XXY」時，人類的認知系統也應該讓讀者能夠輕易地外推理解「*XXY」與「*AAB」有同樣的抽象意義。

方式，那麼人類自然也能合理地為了生存演化出能夠更快讓彼此互相理解、經過特化的先天語言認知功能。語言學的研究也發現人類的確在嬰幼兒時期就已經開始發展環境語言的結構知識，而這樣的快速且特化的語言發展甚至是建立在遠小於大型語言模型的訓練資料量與耗能之上（Fong, 2025）。

四、結語

LLM 發展與人類的語言系統有交集之處，但兩者似乎存在著更多的基礎架構的差異。因此，筆者對「將 LLM 視為代表人類語言系統的理論並研究其語言能力」是否能讓我們更了解人類語言的本質存有高度的懷疑。近年的 BabyLM 挑戰也試著以五歲幼兒的合理語言輸入量（估計約 1,000 萬詞）訓練以 ANN 為基礎的語言模型，而這些模型除了只能完成特定語言任務之外，任務表現也並不意外地缺乏穩定性。據此，筆者認為語言學研究「提出人類語言知識的理論並解釋人類語言表現」的核心任務並不會因為 LLM 的發展而有任何的改變。LLM 是一個以電腦工程和商業應用為本的產物，而不是以解釋人類語言本能為中心的理論架構。在語言學家與 LLM 的核心任務有著根本分岐的情況下，前者若是因為 LLM 在技術與商業層面的成功而對 LLM 在語言學研究中扮演的角色有錯誤認知，才可能真正導致語言研究邊緣化。相反地，基於過往對人類語言知識的研究累積，語言學家或許也能夠改變目前主流 LLM 以無止盡地堆疊算力與擴展訓練集試圖優化模型表現的生態。要達到這個目標，語言學家會比以往更需要跨領域的知識、經驗以及參與。而對筆者來說，若語言學家若能同時具備「扎實的理論語言學背景與實證能力」、「將理論語言學研究成果轉換為實際應用的工程技術知識」和「在跨領域社群分享的熱情」，將會在未來的人工智慧熱潮中有更實質的影響力。

參考文獻

Berent, I. (2013). The phonological mind. *Trends in Cognitive Sciences,* 17(7), 319-327. https://doi.org/10.1016/j.tics.2013.05.004

Berent, I. (2025). Who is afraid of innate knowledge? *Cerebral Cortex,* 35(2), bhaf018. https://doi.org/10.1093/cercor/bhaf018

Bybee, J. L. (2001). *Phonology and Language Use*. Cambridge, UK: Cambridge University Press.

Culbertson, J. (2012). Typological Universals as Reflections of Biased Learning: Evidence from Artificial Language Learning. *Language and Linguistics Compass,* 6(5), 310-329. https://doi.org/10.1002/lnc3.338

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition,* 139, 71-82.

Fong, S. (2025). The Generative enterprise is alive. *Italian Journal of Linguistics,* 36(2), 83-106. https://doi.org/10.26346/1120-2726-239

Gallistel, C. R., A. L. Brown, S. Carey, R. Gelman, & F. C. Keil. (1991). Lessons from animal learning for the study of cognitive development. *The epigenesis of mind,* Chapter 1, ed. by S. G. Carey and R. G. Gelman, 3-36. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Kodner, J., Payne, S., & Heinz, J. (2023). *Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023)* (No. arXiv:2308.03228). arXiv. https://doi.org/10.48550/arXiv.2308.03228

Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex,* 167, 82-85. https://doi.org/10.1016/j.cortex.2023.07.003

Piantadosi, S. T. (2024). Modern large language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett*. Berlin: Language Science Press. Pp.353-414.