大東文化大学大学院中国言語文化学専攻

『中国言語文化学研究』第13号（2024）

語言學與人工智慧

Linguistics and Artificial Intelligence

蔡 維天

Wei-Tien Dylan Tsai

提要 本文討論生成式人工智慧(Generative AI)與生成語法(Generative Grammar)兩套知識體系在語言學、哲學、電腦科學及大腦科學上的同與不同之處，並闡述語法體系在生理界和心理界上的運作條件，及其在人類演化史上的重大意義。這個議題可看成是電腦與人腦的大比拼，對產業界和教育界都有直接且重大的衝擊。首先，我們主張機器學習應「以人為師」，不要自外於人類語言演化的奇蹟。本文以漢語為例，介紹普遍語法(Universal Grammar)的整體規模及核心理念，藉由不同語言間的對比分析來說明語法的通性和個性是如何兼容並蓄，成就了人類今日高度發展的文明。其次會介紹語法做為一個知識體系和符號系統，它的內在數理基礎及其於現實世界的外顯功能。這跟生成式人工智慧的演算模式和模擬機制形成鮮明而有趣的對比，特別是在自然語言處理、資料科學和語言教學方面的應用。最後也希望提醒大眾和學界：即使人類智慧已可模擬到百分之九十九，我們仍應追尋那百分之一的靈光乍現，求真求善，使之臻至完美。

關鍵詞 人工智慧，語言知識，普遍語法、大數據研究、數位人文

Artificial Intelligence, Knowledge of Language, Universal Grammar, Big Data Research, Digital Humanities

目録

1. 如何建構真正的人工智慧?

人類覺得困難或單調的事，機器在明確的規範下能積累經驗案例來類推解決，例如掃地、下圍棋、情資蒐集、圖片辨識、合約判讀等。然而其極限也很清楚：亦即人有多聰明，機器就有多聰明。而許多對人類而言輕而易舉的事，對機器而言卻極其困難，如道德判斷、邏輯推理、理解言外之意及設身處地為人著想(同情心、同理心)等等。我們知道早期人工智慧的發展其實和語言學(尤其是生成語法)息息相關，而所謂自然語言處理(natural language processing; NLP)可視為人工智慧(artificial intelligence; AI)的一部分，就如同語言是人類智能的一部分。如果我們將人工智慧定義為一個人類大腦智能的模擬器，那麼自然語言處理便提供了我們跟這部模擬器之間溝通對話的管道。

歸根究底，Artificial Intelligence (AI)的關鍵字是*artificial*，即使是「機器學習」，其實也是由人為的演算法所主導。因此我們主張人工智慧仍應以人為師，承襲人性倫理的考量，以免因失控遭到反噬，甚至成為少數利益團體掌控的工具，造成文明社會不可逆轉的損傷。這也正是知名天文物理學家霍金(Stephen Hawking)在BBC的訪談中提出下列警訊的原因：

The development of full artificial intelligence could spell the end of the human race.
人工智慧全盤的成熟發展可能意味著人類的終結。

其實諸如此類對AI未來的擔憂在文學和影劇中屢見不鮮：科幻小說大師艾西莫夫(Isaac Asimov)便曾為機器人的發展立下了三大法則(Three Laws of Robotics)：[1]

I. The First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
第一法則：機器人不能傷害人類，也不能見死不救。

II. The Second Law: A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
第二法則：機器人必須聽命於人類，除非這道命令與第一法則相抵觸。

III. The Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
第三法則：只要不抵觸第一法則及第二法則，機器人必須全力自保。

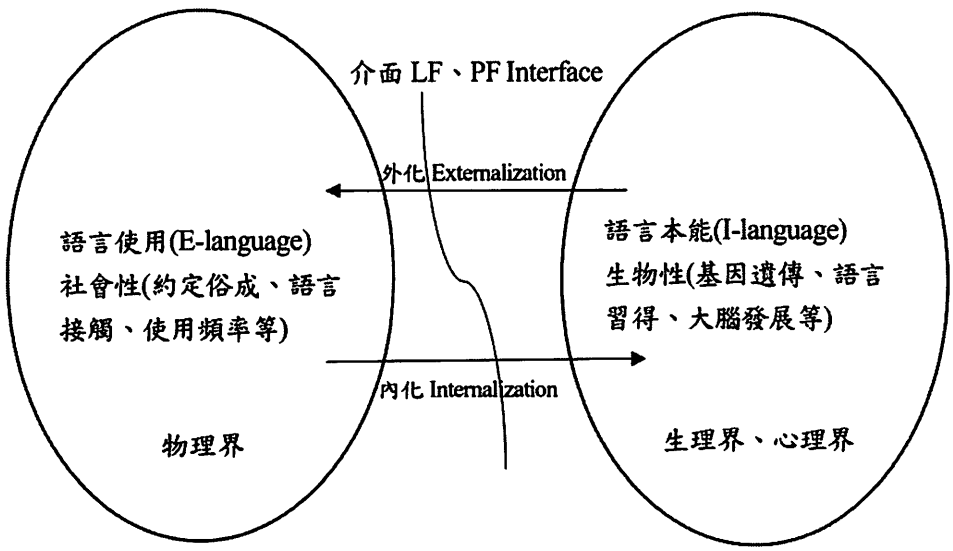
大導演庫柏力克(Stanley Kubrick)的經典巨作 *2001: A Space Odyssey* (中譯《2001太空漫遊》)中太空船的人工智慧 HAL 9000 便因任務邏輯上的衝突對成員起了殺意並開始付諸實行，過程中明顯將第二法則及第三法則置於第一法則之上。而1984年的影片*The Terminator* (中譯《魔鬼終結者》)及其後續

[1] 語出1942年艾西莫夫所著小說*Runaround*，後收錄於1950年文集 *I, Robot*。

系列作品更是直接演繹AI失控的可能性，劇中先進人工智慧SkyNet產生自我意識，並判斷人類才是地球最大的威脅，因而發動核戰意欲毀滅其創造者。

當然，這些所謂未來學(futurology)中的場景離現實還有一段距離，我們現在首先要面對的是人工智慧發展背後的利益集團，他們龐大的投資將如何回收？人類的終極福祉是否能像機器人法則一樣放在首位，或是隨時都可以在收益考量下成為犧牲品？也因此科技巨擘馬斯克(Elon Musk)認為AI可能是人類生存最大的威脅之一(*one of the biggest threats to humanity*)；甚至連Open AI公司自身的研究團隊都提出類似警訊，揭露盲目追求強大AI的危險性。

對一個追求真理的學者而言，更重要的問題是語法研究最後的目標，究竟應該放在徹底解析人類智識體系的真相，還是複製出一個好看好用的仿品？或許現實的答案會在兩者之間擺盪，但在億萬年演化的奇蹟之前，我們是否應該更謙卑一點，真正「深度學習」語言基因的奧秘？這套經過時間淬鍊的語言知識(knowledge of language)其實原本就有由外而內的面向，如索緒爾(Ferdinand de Saussure)提出的約定俗成原則(arbitrariness)；也有由內而外的面向，如人類孩童語言習得(language acquisition)的進程。我們可以通過語法介面(grammar interface)的概念來理解其間的關聯，進而釐清喬姆斯基(Noam Chomsky)所謂「外延語言(E-language, extensional language))」和「內涵語言(I-language, intensional language)」之間的互動關係，如下圖所示：



中國儒家嘗謂「內聖外王」，其實自然語言也可以這麼看：內涵語言是基於普遍語法(Universal Grammar, UG)而得來的一套內在語言知識(innate knowledge)，或者亦可稱之為語言本能(language instinct)。這套知識具有個體性，是人類心智系統的一部份。而外延語言是約定俗成的外在語言行為，具有高度的

社會性和文化性，跟現實世界中語言使用(language use)有密不可分的關係。

2. 人工智慧的前世今生

前面提到，資訊科學和人工智慧早期的發展跟生成語法(generative grammar)有相當密切的關連。而近年來人工智慧則與資料科學(data science)、影像辨識(Image recognition)和機器翻譯(machine translation)的研究相結合，建構出大型語言模型(Large Language Model, LLM)，在運算速度和實用性上都有飛躍性的成長，其所依賴的要素有三：

I. 一部運算力強大的機器
II. 巨量的數據庫
III. 一套有效率的演算法，如梯度下降(gradient descent)、類神經網路(artificial neural network)等

機器耐得住反覆單調的運算，因此往往比人類更能求出最優解，也就是將預測和現實的差距最小化，來達到訓練的目的。有趣的是，此三者沒有一個是有智慧的，卻能模擬出具有人類智能的表象，甚至讓測試者懷疑其已具備自我意識(self-consciousness)。

時人有言：ChatGPT是人工智慧的里程碑，更是一道分水嶺。從基本為用戶解答疑難，躍升到自主思考及創作內容的境界。先不管這是商業操作，還是過度樂觀，但是通過巨量數據的蒐集，用戶們積極參與*trial & error*的回饋過程，機器翻譯和機器學習的確有了長足的發展，並在實用化的面向上突飛猛進，商機無限。而最近也有一種看法是以「自動化前沿(frontier of automation)」這個量化概念來衡量「通用人工智慧(AGI)」的發展：亦即分工自動化的複雜度愈大，其任務整體的達成度也愈高，甚至在不久之將來就會超過人類解決問題的能耐。但有得必有失，面臨失業衝擊的族群就包括了企業公司的服務端、低階碼農、翻譯工作者、文案工作者、視覺藝術家。另一方面，發展人工智慧所需的數據蒐集則面臨侵犯隱私權(如行跡追蹤、人臉辨識、消費行為預測等)等種種問題。而無人機(drone)的交戰規則(rules of engagement) 更是引發了強烈的倫理爭議：若是全面自動化勢必會違反機器人三大法則，其使用規範自然成了文明國家關心的重大議題。

如眾所皆知，人工智慧研究的先驅圖靈(Alan Turing)提出了圖靈測試(Turing Test)，其本質是基於外延語言的一種判準：如果受試者無法分辨出與其對話的是人還是機器，那就具備了人類智能的水平。這個測試簡單粗暴卻非常有效，長年以來深受大家所喜愛，甚至常有國際大賽讓有志者一分高下，看看誰的程式最能把人唬住。[2]也因此我們可以把ChatGPT的成熟發展視為自然語言跟人工語言的對決，

[2] 圖靈測試也有貼合時代脈動的新解：例如科幻電影 *Ex Machina* 揭示了圖靈測試的新頁：即使受試者知道對方是機器，是否還能產生同理心，將之視為人類來對待？電影標題的典故出 *deus ex machina* "god out of the machine"，為希臘戲劇用語，由演員扮演神名來解決一場不可解決的衝突困局。這其實也暗示了圖靈測試可能無解，模擬究竟是模擬，對了解人性、智能未必會有幫助。

喬姆斯基甚至將其定性為高科技剽竊(high-tech plagiarism)的產物。當然，ChatGPT雖然已經超越早期ELIZA採用模式匹配(pattern matching)的對話策略，採用所謂類神經網路的演算法，但仍脫不出線性—概率的計算語言模型，無法完全掌握自然語言最重要的兩個數理特質，亦即合併(Merge)與遞迴(Recursion)。喬氏的警語猶言在耳，做為人文社會學界的中堅力量，我們是不是也該做個「深度學習」，重新思考自然語言的本質？

換句話說，ChatGPT即使再好用、再好玩，從探索人類大腦智能的角度來看，也只是極其表象的東施效顰。這也是圖靈測試這種「結果決定論」的極限和盲點，其簡單粗暴也正是其吸引人之處：在能真正拆解大腦這個黑盒子之前，圖靈測試也不失為一個有效的語言工程判準。我們認為要擺脫掉工人智慧的桎梏，文人智慧更應該從一個批判的角度出發，梳理AI和人類心智的同與不同之處，建立辨別真品與仿品的知識和流程：一方面讓物盡其用，將AI應用導向正途；另一方面則應持續在物種演化、心理認知及道德判斷的面向上探索人性，就像瞭解心臟的構造一樣來瞭解人腦的構造，建立能真正能反映出語言知識模型。

為了求知求真，即使人類智能已經被模擬到百分之九十九，我們仍應尋求那百分之一的靈光乍現。日本有一部名聞國際的科幻動畫電影「攻殼機動隊」，深刻描繪了生化人(cyborg)面臨的倫理難題(如人工移植的記憶算不算真正的記憶？而擁有假造記憶的人又該如何看待自己？)，其英文片名是*Ghost in the Shell*：至於這個“ghost”是鬼魅還是聖靈，端看人類自己的做為以及公民組織的監督，才能防止利益集團無限上綱的掌控與濫用。

3. 內涵語言的基本運作機制

相對所謂生成式AI (Generative AI)從外延語言出發的巨量操作，生成語法(Generative Grammar)則更關心語言做為人類本能所具有的數理性質，在此我們介紹兩個最重要的概念：其一為「合併(Merge)」，亦即把兩個單元合併為一個組合(constituent)；另一則為「遞迴 (recursion)」，亦稱「遞歸」，其定義為一個組合與另一個組合合併後仍是一個合法的組合，最常見的例子就是語法結構中詞組和詞組結合後成為另一個更大的詞組。後者很可能就是區別人類語言和其它物種溝通系統的關鍵：此項機制可以生成無限長的句子及無限多的句子，提供了語言創新性的數理基礎(請參見Hauser, Chomsky & Fitch 2002)。這也讓我們想起下面這句出自《老子》的千古名言：

道生一，一生二，二生三，三生萬物。

我們可以用李白膾炙人口的一首唐詩來說明上述理念，下面是用台灣閩南語羅馬字標注的《靜夜思》：

床前明月光，

tshong^{5} tsian^{5} bing^{5} guat^{8} kong^{1}

疑是地上霜，

gi^{5} si^{7} te^{7} siong^{7} song^{1}

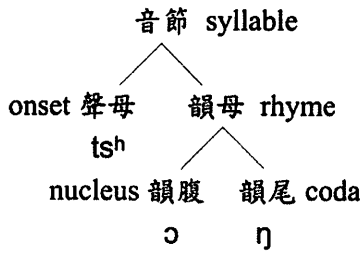
舉頭望明月，

ku^{2} thio^{5} bong^{7} bing^{5} guat^{8}

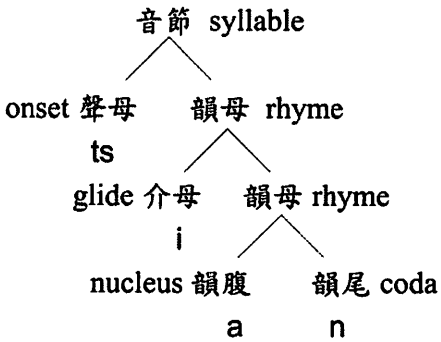
低頭思故鄉。

te^{1} thio^{5} su^{1} koo^{3} hiong^{1}

從音韻學來分析「床」字的音節結構，我們就會發現從音段(segment)開始就是疊床架屋，充分運用合併的生成機制：

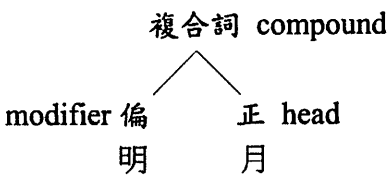


接下來還可以採用遞迴機制來增加一層韻母的組合，以追加介音，如下面「前」的音節結構所示：

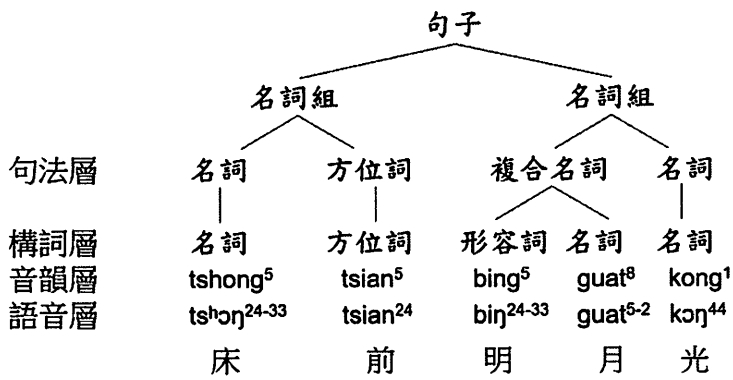


就構詞學而言，一個詞如「床頭」可以分析為詞根(root)和詞綴(affix)的組合：

而複合詞(compound)如「明月」則是以古漢語語法為根底的偏正結構(modifier-head structure；參見Chao 1968、梅廣2002)，由修飾語「明」和中心語「月」合併而成：

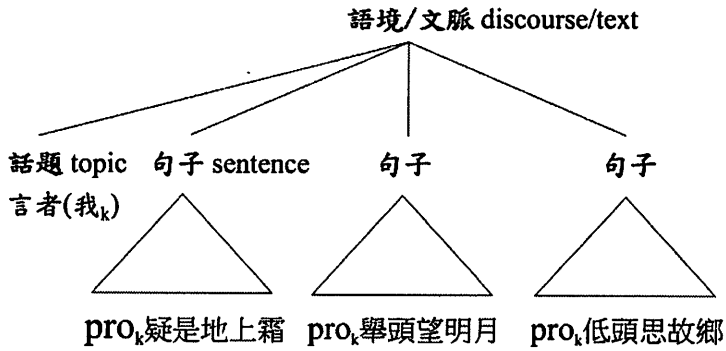


及至句法學，遞迴機制則更被大量使用，如下圖所示：



最後到篇章之學的層次，其語境/文脈(discourse/text)仍可用邏輯上的組合原則(compositionality)來解析其層系結構，合併、遞迴仍扮演著重要的角色。請參考下面對《靜夜思》後三句的話題鏈分析(topic chain analysis；請參見Tsao 1990)：[3]

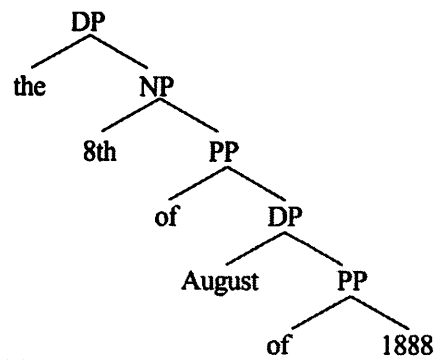
[3] 圖中 *pro* 表空代詞之意。此即生成語法中所謂的主語代詞失落(pro-drop)現象，是漢語一個很重要的類型特色。



Hauser, Chomsky & Fitch (2002)將上述數理機制視為人類語言運算系統的核心，可說是億萬年演化的結晶。普遍語法(UG)做為語言基因遺傳通性的起源，其定義是「人類孩童語言發展的初始狀態」。因此把UG想像成一套教科書上的語法規則其實並不太恰當，更好的比喻應是將其比做一株幼苗：因此我們不說小孩學會了母語，而是說他們的語法長大了；根苗雖小，卻已經有了大樹的藍圖。然而這又引發了另一個問題：如果人類孩童語言發展的起點一致，那為什麼長大成人後，每個人的語言卻都是獨一無二的？關鍵就在於外在環境教養(nurture)的影響。換句話說，孩童從呱呱墜地以來的所見所聞就是養分，讓他們在有限的選擇下設定參數(parameters，如英語的VO還是日語的OV)，其語言能力得以迅速成長，兩三歲便能將母語講得呱呱叫，這也正是語言習得的奧義及其跨界研究的迷人之處。

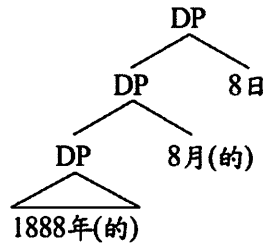
關於參數的設定，我們比對中英文日期的寫法就會看得很清楚：「1888年8月8日」的英文可寫成"8th August, 1888"，拆解開來其實就是"the 8th of August of 1888"，這其實是一種中心語在前(head-first)而右向分枝(right-branching)的名詞組結構，如下面樹圖所示：

(1)



而中文的名詞組則跟日文相近，屬中心語在後(head-final)而左向分枝(left-branching)的結構，因此相當於「1988年的8月的8日」。這點比對下圖便可看得很清楚：

(2)



更有趣的是，兩者在參數設定上的區別也充分反映出其地址寫法之上：如“101, Sec.2, Kuang-Fu Rd., Hsinchu, Taiwan”亦即“101 of Section 2 of Kuang-Fu Road of Hsinchu of Taiwan”，屬中心語在前的名詞組結構。而中式寫法則剛好倒過來，寫成「臺灣新竹市光復路二段一百零一號」，拆開成為就成了「臺灣的新竹市的光復路的二段的一百零一號」，可說是英式寫法的鏡像(mirror image)，亦即中心語在後的結構，這點跟日文的名詞組也非常類似。

4. 語言知識的內外之別

外延語言經常以 *trial and error* 的方式來累積經驗，因此和語言使用(language use)高度相關。科技界透過大數據訓練來通過圖靈測試的嘗試，近年來已有長足發展，如Siri、Alexa、ChatGPT等產品都是絕佳例證。相對而言，內涵語言這套知識體系植基於大腦，是億萬年來演化的結晶。事實上，對於人腦功能及其模組互動，我們大多只知其然，不知其所以然。因此普遍語法仍是至今學界最明確、最具解釋力的理論模型之一。

要瞭解人類智能的奧秘勢必要先解碼其語言體系所涵養的知識。而這套知識可說是 *know without knowing*，就如會開車的人未必知道車子的內部構造原理，說話的人也未必知道語法的組織和結構，卻都可以講得很好。但也因此很多人都有一種似是而非的觀念，覺得自己懂漢語，就懂漢語語法。一旦學者開始用科學方法來研究人類語言，才發現原來我們知道的這麼多，卻又知道的這麼少。而其策略則不出異中求同和同中求異：前者以找出放諸四海而皆準的原則系統為終極目標；後者則尋求萬變不離其宗的參數設定，以解釋語言在表象上的紊亂與分歧。

另一方面，我們也可以說語言的發展過程是 *learn without learning*：小孩子學母語其實不是用學的，是語言基因在外在經驗的滋養下慢慢成長；這跟成人後的記誦式學習完全是兩回事。因此內延語言的習得對機器來講是很困難的，而外延語言則可靠勤學而致之(如同muscle memory)，而這也正是機器學習和二語習得(second language acquisition)共通之處。

誠然，人工智慧研發是新世代的顯學，是菁英人才和巨量投資的匯聚之所，但我們在窮盡種種訓

練機器的演算法之餘，不妨也回頭看看人類語言的本源活水：竊以為機器學習跟語言習得、古文教育頗有相通之處，問題在於如何結合語法理論和概率統計在關鍵處植入語言知識，向億萬年來人腦進化的奇蹟致上敬意。以人為師，相信不會空手而回的。

5. 腦科學與人工智慧

腦科學(brain science)和語言學的跨界研究稱為腦神經語言學(neurolinguistics)，有兩條主要的思路和理論模型：其一是基於普遍語法的理念，認為孩童學習語言是有所本的：語法從基因到成形只受到外界經驗有限的刺激，卻能由內而外建構與呈現對這個世界的理解，並體現於我們的外在行為(如溝通交際、寫作修辭等)。這個理論模型在心理語言學、計算機語言學及生物語言學界(biolinguistics)有其一定的影響力。

另一個研究趨向稱為連結理論(associative theory)，認為語法並不是一套與生俱來的知識或運算系統，而是由外而內經由數據的蒐集分析來學習規律，辨識範疇。也因如此，語言知識是可以因外界資料的變動而被改寫的。這條思路廣為資訊科學界所接受，如大家耳熟能詳的類神經網路、深度學習(deep learning)，並廣泛應用在圖形辨識、自然語言處理、自動駕駛、網路遊戲博弈等機器學習(machine learning)的項目上。

Moro (2010)一書以巴別塔的分際(The Boundaries of Babel)為隱喻來談自然語言中先天生理基礎和後天文化教養的互動，從各種不同的角度來深度探索內涵語言及外延語言的分野。他提到語言習得若是由內而外，則一定有其先天原則系統的限制，應該會有所為而有所不為，有些語句是永遠不可能出現的。而由外而內來建構語法則不然，其原則是所謂的約定俗成，就如同維根斯坦(Ludwig Wittgenstein)的「語言遊戲觀」一般，依社交情境脈絡而來的遊戲規則應該是沒有先天極限的。Moro (2016)則進一步「不可能的語言(impossible language)」這個概念；其研究團隊的開腦實驗顯示，人類語言的確有所為而有所不為：詞組語法在布氏區(Broca's area)腦血流的激活程度上遠比線性語法(亦即不可能的語言)來的高。

時至今日，已經有腦科學家可以從腦波解讀出足以自動合成語音的指令，間接證明了PF介面的存在，並造福有發音障礙的病人(請參見Anumanchipalli, Chartier & Chang 2019)。更有趣的是，還有研究在更大的實證基礎上訓練AI將腦波轉譯成語句，使其過程更快、更有效率(請參見Makin, Moses & Chang 2020)。從這裡也可以看出，我們還能透過人機介面將語言知識呈現出來，無論是由內而外，還是由外而內，都能更進一步參透心理界、生理界和物理界錯綜複雜的互動關係。

6. 語法研究、語言教學和人工智慧

第二語言習得(Second Language Acquisition，簡稱二語習得)的研究顯示，一個人學習外語的過程其實並非從頭開始，而是會受自己母語的影響，甚至可以說是以母語語法為基礎來學習外語。正如我們所熟知的，母語對學習外語會形成一種干擾作用，造成許多發音、用字和文法上的錯誤(例如將CD唸成「西滴」)。然而我們往往忽略掉一個事實，那就是語言之間有許多共通之處：母語其實是助力，而非阻力。為我們學習外語提供了極大的助益。以此觀之，如果我們要訓練機器說話寫作，與其從頭開始，還不如把它當做外國人來教，先餵一些簡單明瞭的語法規則，再求跟數據的演算結合，達到事半功倍的效果。

除了外語翻譯和教學方面的挑戰，新一代的人工智慧還需要面對外延語言中常見的語言接觸(language contact)、語言變遷(language change)等現象。新加坡英語(Singlish)就是一個很好的例子，我們先檢視下面這個混合語(creole)的句式：*Google Translate*明顯是照字面硬翻，*ChatGPT 3.5*則翻得相當到位，對其特殊情態(modality)用法做了精確的詮釋：

(3) You die die must try!
Google Translate: "你死死一定要試試！"
ChatGPT 3.5: "你非試不可！"、"你死都要試試！"

此外，對漢語中較為特殊的疑問詞條件句(*wh*-conditionals)，*Google Translate*完全不知所云；相較之下，*ChatGPT 3.5*表現亮眼，採用英文的自由關係句(free relatives)做了適切的表達：

(4) 你愛怎麼樣，怎麼樣。
Google Translate: "How do you love, how do you love."
ChatGPT 3.5: "You can do whatever you want."

上述兩個案例皆顯示OpenAI演算法結合巨量資料耙梳的確具有其優勢，朝著翻譯的終極目標「信」、「達」、「雅」大步前行。

在語言演化的面向上，我們注意到現代漢語有一種「及物化(transitivization)」的趨勢：不及物(或是半及物)的複合動詞產生了及物用法：原本在動詞的非核心論元(non-core argument)，卻放到動詞後典型的賓語位置(參見蔡維天2016a, 2017)。比較成熟的案例如從(5a)轉化出(5b)，從(6a)轉化出(6b)：

(5) a. 我[對這件事]很在意。
b. 我很在意[這件事]。

(6) a. 我[對這個人]很關心。
b. 我很關心[這個人]。

漢語做為一個典型的孤立語(isolating language)，一直在語言類型學(linguistic typology)上佔有獨特的地位：所謂「孤立」是指構詞時少用黏著形式(bound forms)，一個詞基本上由一個詞素組成。另一個相互相生的特色則屬其強勢的分析性(robust analyticity)，亦即構句較少使用屈折詞素(inflectional morphemes)，表達語意多用組合方式而非單詞。這點在上古漢語到中古漢語的演化進程上尤為顯著，其句法類型由綜合(synthetic)轉為分析(analytic)，如文言「壯士死知己」到白話中就變成「壯士為知己而死」；文言「孔明哭周瑜」到白話中就變成「孔明為周瑜哭泣」(參見黃正德、柳娜2014；蔡維天2016b)。

有趣的是，這個進程近年來有大幅反轉的趨勢，主要出現在兩種文體之中：其一是媒體標題，如「伊朗跟沙國斷交」變成「伊朗斷交沙國」，又如「蕭亞軒跟百億男友分手」變成「蕭亞軒分手百億男友」。另一個則是口語文體，有下列Youtube的自然語料為證：

(7) 你是不是想要仙人跳我？ ≈ 你是不是想要對我(做)仙人跳？
(小明星大跟班 2021.03.04, 46:32)

(8) 他們有默契地宣布停止捐款共和黨 ≈ 宣布停止給共和黨捐款
(文茜的世界財經周報 2021.01.17, 7:57)

(9) 女星下跪每個乘客 ≈ 女星對每個乘客下跪
(CTWANT, 2021.02.07)

(10) 我不熟這個年輕人 ≈ 我跟這個年輕人不熟
(中時新聞網, 2021.01.16)

問題是我們該如何以科學方法來解析這些尚在發展中的趨勢，又該如何提出合理的推測，做為今後驗證調查的可靠指標？事實上，現今語法研究正面臨著兩難的困境：在實證層次上，學界時興用語料庫及搜尋引擎(如Google)來支持某些特定論述；但另一方面，學界、業界的研究則顯示人類的語言能力和知識絕非特定的演算法可以模擬，其簡約的數理機制及普遍原則(如合併、遞迴)具有不可羈握的原創力。我們認為問題的關鍵在於驗證的範圍太窄，採樣數量也遠遠不足，以致無法切實掌握真象，這也正是語法研究可向資料科學(data science)借鏡之處。做為此類跨界合作的先導研究，台灣的清華大學團隊針對及物化的議題，比較兩個不同年代的新聞資料庫，得出下列兩點結論(參見蔡維天、楊磐

瑜、陳映竹、陳志杰、張俊盛2022)：

其一、新聞標題比內文更傾向使用及物句構。

其二、年代較近的新聞資料比較久遠的資料更傾向使用及物句構。

如此一來，我們便可結合語法理論和自然語言處理等領域的研發成果，在基礎研究和實證應用兩方面取得平衡，進而達成相輔相成、相互發明的雙贏局面。

7. 結語

從人工智慧的角度來觀測語言現象並印証於語法理論，其潛在優勢與能量令人期待：首先透過巨量資料的蒐集、分析，我們能提供語言演變的即時資訊，呈現語言的使用現狀，並預測其未來發展趨勢。具體而言，藉由大數據工具在時間及空間二個向度上蒐集資料，我們不但可以逐年驗証漢語類型是否正由分析性轉回綜合性的老路，還可以經由標定IP位置研究語言、方言在地理上歧異、接觸與混合。假以時日，相信學界便可開始打通從內涵語言映射至外延語言的任督二脈，從根本上釐清語言的生物性和社會性之間錯綜複雜的關係，讓文人智慧和工人智慧一起創造出真正的人工智慧。

參考文獻

Anumanchipalli, Gopala, Josh Chartier, and Edward Chang. 2019. Speech Synthesis from Neural Decoding of Spoken Sentences. *Nature* 568: 493–498.

Chao, Yuen-Ren 趙元任. 1968. *A Grammar of Spoken Chinese*. Berkely: University of California Press.

Hauser, Marc, Noam Chomsky, and Tecumseh T. Fitch. 2002. The Faculty of Language: What is it, Who has it, and How did it evolve? *Science* 298(5598): 1569-1579.

Makin, Joseph, David Moses, and Edward Chang. 2020. Machine Translation of Cortical Activity to Text with an Encoder-decoder Framework. *Nature Neuroscience* 23: 575-582.

Moro, Andrea. 2008. *The Boundaries of Babel*. Cambridge, Massachusetts: The MIT Press.

Moro, Andrea. 2016. *Impossible Languages*. Cambridge, Massachusetts: The MIT Press.

Tsao, Feng-Fu 曹逢甫. 1990. *Sentence and Clause Structure in Chinese: A Functional Perspective*. Taipei: Student Book Company.

蔡維天. 2016a.〈論漢語內、外輕動詞的分佈與詮釋〉，《語言科學》，第15卷第4期(總第83期)，362-376頁。

蔡維天. 2016b.〈別鬧了，余光中先生！一從生成語法的角度檢視語言癌之生成〉，《語言癌不癌？語言學家的看法》，51-78 頁，台北：聯經出版公司。

蔡維天. 2017.〈及物化、施用結構與輕動詞分析〉，《現代中國語研究》，第19期，1-13頁。
蔡維天、楊聲瑜、陳映竹、陳志杰、張俊盛. 2022.〈漢語及物化的大數據研究〉，《台灣語言學期刊》，第20卷，第1期，1-27頁。
黃正德、柳娜. 2014.〈新興非典型被動式"被XX"的句法與語義結構〉，《語言科學》13.5: 225-241。
梅廣. 2002.〈迎接一個考證學和語言學結合的漢語語法史研究新局面〉，《漢語的歷史發展》， 何大安編。台北：中央研究院。