Taiwan Journal of Linguistics
Vol. 20.1, 1-27, 2022
DOI: 10.6519/TJL.202201_20(1).0001

漢語及物化的大數據研究*

蔡維天[1]、楊謦瑜[2]、陳映竹[1]、陳志杰[1]、張俊盛[1]
[1]國立清華大學、[2]國立中興大學

本文從資料科學的角度來考察漢語中一個新興的現象「及物化」：亦即動前介詞組轉化為動後的直接賓語（如「為人民服務」轉為「服務人民」）。此現象仍處於變動之中，可視為一種復古的趨勢（如「士為知己而死」就有相應的為動式「君子死知己」），可能會逐步消亡，也可能像「語言癌」一般引發爆炸性的發展（參見何萬順等 2016）。因此我們需要結合資料科學和語法理論來為其問診把脈，不但要總結先前的演化歷程，更能預測未來的發展趨勢。此外，及物化現象只在幾類文體中展現高度的能產性，這也讓我們有機會一窺其使用上的考量及背後機制，並印證於新聞語料庫的統計數據之上。

關鍵字：及物化、漢語句法、輕動詞、資料科學

* 作者在此感謝 "自然語言與語音處理研討會（ROCLING）" 與會學者的建議與指正，以及國立清華大學自然語言處理實驗室的大力協助。此外，《台灣語言學期刊》編輯部及審稿專家的評論和建議也對整體修訂有莫大助益。本文寫作期間獲得科技部計畫（MOST 108-2633-M-007-001, MOST 109-2639-M-007-001, MOST 106-2410-H-007-030-MY3）的經費資助，在此一並致謝。

1. 緒論：何謂及物化？

在歷史的長河中，語言的變遷無疑是學者們考察關注的重點之一：其原因即在於漢語有大量的文本典籍和歷史記錄（如災荒、遷徙、戰爭、駐軍及地緣政治等）可供查對，同時還有從唐宋以下各朝各代建立起的韻書撰寫傳統，反映出多層次、多方言的動態音韻體系，豐富了我們探究漢語歷史演化的工具箱，讓專家學者得以一窺中古漢語和上古漢語的風貌，也使重建祖語（proto-language）成為可能。

而現代語言學的發展也正處在一個關鍵期：語言做為「人之所以為人」的生物本能是如何跟種種外緣因素互動（如使用、接觸、混合等），進而驅動了演化的歷程。這些困難複雜的議題都需要有嚴謹的田野、實驗和大數據研究做為工具，才能抽絲剝繭，澄清問題的本質，並提出明確的解決之道。

根據以上理念，我們將探索重點聚焦在一個近年來興起的語法現象，可姑且稱之為「及物化」（transitivization）。漢語文獻中其實已經有了相當深入的觀察和後續討論：齊滬揚（2000）即指出「漢語述賓結構是一種優勢結構，許多原非述賓結構有向述賓結構靠攏的趨勢這樣使得一些原先不能帶賓語的動詞也逐漸可以帶賓語了，及物動詞的數量呈擴大趨勢。」他舉了以下幾個淺顯易懂的例子，可作為參考：

(1) a. 他常[為人民]服務。
b. 他常服務[人民]。

(2) a. 他常[用毛筆]寫字。
b. 他常寫[毛筆]字。

(3) a. 我們[在北京]相見。
b. 我們相見[北京]。

如(1a)中動前的由「為」引介的受惠者「人民」在(1b)中變成了動後的直接賓語；又如(2a)中動前的由「用」引介的工具「毛筆」在(2b)中變成了動後的直接賓語；[1] 最後(3a)中動前的由「在」引介的地點「北京」在(3b)中變成了動後的直接賓語。

這種及物化的現象其實在古漢語就極為發達，如東晉陶淵明〈詠荊軻〉的名句「君子死知己」即「君子為知己而死」的意思，是為動式很好的例子。事實上，馮勝利（2005）指出先秦漢語中「小」有兩種及物用法：如《孟子》中「孔子登東山而小魯，登泰山而小天下」即屬意動式，表「以魯為小」、「以天下為小」之意；而「匠人斫而小之則王怒」則屬使動式，表「斫而使之小」的意思。這些句法機制不但在前述及物化現象中重現，近年來還在被動式中「借屍還魂」，如「被消失」有「被致使消失」的使動式用法，也有「被當成消失」的意動式用法（詳見黃正德、柳娜 2014）。如此看來，語言演化的趨勢

[1] 此處評審學者指出「寫毛筆字」未必是及物化的例子，因為「寫字」本身已是動賓結構。這點在蔡維天（2017）中的形式分析中已有回應：亦即「寫毛筆字」在語意上和「用毛筆寫字」相對應，但在句法上確需要經過動詞移位（verb movement）來調整詞序。更具體一些來說，漢語依其類型學上設定允許隱性輕動詞，因此「寫毛筆字」的句法基底結構其實是「WITH 毛筆寫字」，WITH 則理解為聽不到的「用」，但和「用」不同之處在於可吸引主要動詞「寫」上移，形成「[寫+WITH]毛筆字」的表層結構。無論是「寫鋼筆字」、「寫鉛筆字」都是一個道理。此外，審查意見也提到「服務」的「務」有「務農」、「務實」的及物用法，因此或許不是及物化的例子。此處需要注意兩點：其一，「務農」、「務實」正是古漢語及物化在複合詞中存留下來的現象，亦即「以農為務」、「以實為務」的意思。其二，「服」與「務」結合起來的並列式已進入另一個階段的新發展，不像詞組有清晰透明的組合性（compositionality），這也是漢語複合詞的特色（請參見梅廣 2003）。

也同樣是合久必分、分久必合，其類型常在綜合性和分析性之間擺盪。

蔡維天（2017）則注意到上述及物化現象在某些文體或地區特別顯著：比如說臺灣媒體下簡短標題時就常採用此種及物句構，表對象關係的例子如(4a-d)，可分別轉寫為(5a-d)，原本動後的直接賓語放到了動前介詞組的位置：

(4) a. 馬林魚仍甜蜜復仇[紅雀]。
b. 張艾亞生日告白[許孟哲]。
c. 脫歐死鬥！保守黨叛變[強生]，英國國會解散倒數。
d. 法院嗆聲[王寶強]。

(5) a. 馬林魚仍[對紅雀]甜蜜復仇。
b. 張艾亞生日[對許孟哲]告白。
c. 脫歐死鬥！保守黨[對強生]叛變，英國國會解散倒數。
d. 法院[對王寶強]嗆聲。

表示與同（comitative）關係的例子如(6a-e)，可分別轉寫為(7a-e)，亦即動前不及物用法的例子：

(6) a. 伊朗斷交[沙國]。
b. 蔡康永牽手[小 S]，…
c. 蕭亞軒分手[百億男友]。
d. 秦凱求婚[何姿]。
e. 海盜升上王維中，碰面[陳偉殷] …

(7) a. 伊朗[跟沙國]斷交。
b. 蔡康永[跟小 S]牽手，…
c. 蕭亞軒[跟百億男友]分手。
d. 秦凱[跟何姿]求婚。
e. 海盜升上王維中，[跟陳偉殷]碰面 …

最後是表達蒙受關係的(8a,b)，可分別用動前的「給」、「把」轉寫為(9a,b)：

(8) a. 法陸空罷工添亂[歐國杯]。
b. 美國聯邦法官打臉[川普]。

(9) a. 法陸空罷工[給歐國杯]添亂。
b. 美國聯邦法官[把川普]打臉。

相對而言，這種新興的及物化在敍事文本中則較少見。更有趣的是，這類用法在口語中反而能產性頗高，而且愈來愈普遍：在動後賓語位置引介對象論元的有(10a-d)，可分別改寫為(11a-d)：

(10) a. 有沒有人發問[你]？
b. 他常輕挑[女生]。
c. 你別刻薄[人家]。
d. 但我絕對一定要來靠北[她]。

(11) a. 有沒有人[對你]發問？
b. 他常[對女生]輕挑。
c. 你別[對人家]刻薄。
d. 但我絕對一定要來[對她]靠北。

此外，引介受惠者或原因論元當直接賓語的有(12a-d)，可分別改寫為(13a-d)中的不及物用法（也有人稱為半及物）：

(12) a. 報紙一直吹牛[他]。
b. 薇如緊張[我]，…
c. 他很傷心[這件事]。
d. 他很高興[這件事]。

(13) a. 報紙一直[為他]吹牛。
b. 薇如[為我]緊張，…
c. 他[為這件事]很傷心。
d. 他[為這件事]很高興。

最後，引介蒙受論元當直接賓語的則是(14a,b) ，可分別改寫為(15a,b)：

(14) a. 他還想要索賠[對方]。
b. 你怎麼可以放鳥[人家]？

(15) a. 他還想要[跟對方]索賠。
b. 你怎麼可以[把人家]放鳥？

此處文體造成的差異非常有趣：正如兩位評審學者所言，標題有字數限制，必須簡明醒目，體裁上需要配合題材，講究的是語不驚人死不休。但這並不代表這種語言使用上的考量和前述及物化的語法手段有所扞格，兩者其實互為表裡：及物化是內延語言（intensional language）的句法機制，基於其類型特色而來（亦即允許隱性輕動詞（implicit light verbs），參見註解 1）；而標題設計則以及物化縮減字數並改動語序，造成極

其鮮明的「吸睛」效果，這正是外延語言（extensional language）成就的事功，在新聞傳播的應用上頗具意義。另一方面，口語中及物化也多出現在賓語為代詞的句子，這顯示其中很可能還有韻律上的考量，因為代詞常會因虛化而失去音韻上的地位，進而驅動了語法上的變動。這些因素都是非常值得深究的議題，但由於可供運算分析的語料仍然有限，本文仍將重點放在及物化在資料科學上的驗證。而文體及韻律方面的議題則需等待未來研究條件成熟（如相關語料庫的取得及建構、語音實驗面向上的跨界合作），才能作更深入、更為切題的探討。

2. 研究方法

我們進一步使用語料庫驗證上述漢語及物化演變的趨勢（參閱蔡維天 2017），考量及物化可能因年代與語體而有不同的變化，本研究選用不同年代的新聞資料作為研究標的以觀察及物化在歷時方面的演變，並區分標題與內文以利考察語體對及物化的影響。我們分別將這三份資料建構成以依存關係（Universal dependency）（Nivre et al., 2015, 2016）為框架的語料庫，並抽取所需資料做統計，方法步驟如下：

(16) 及物化統計方法：
- a. 資料前處理：標點符號正規化、斷句、斷詞、詞性分析、依存關係剖析（Universal dependency）。
- b. 從依存關係剖析結果抽取及物結構與不及物結構。
- c. 篩選抽取結果：目標動詞、介系詞、及物結構與不及物結構的轉換。

本節將介紹使用的資料、進行的預處理與運用的剖析器（Parser）、結構抽取方式、抽取結果篩選方式。

2.1 資料介紹

我們將語言資料協會（Linguistic Data Consortium） 發行的中文十億字資料集（Tagged Chinese Gigaword）[2] 的標題與內文分開儲存為兩份資料庫，如(17a,b)，涵蓋的新聞來源有：中央通訊社、新華通訊社與聯合早報，資料年份為 1991 年到 2004 年，約 181 萬 6 千多則。第三份資料，如(17c)，為 2004 年到 2017 年間的新聞內文，主要新聞來源為：聯合報、聯合晚報、經濟日報，約有 177 萬多則新聞。[3]

[2] https://catalog.ldc.upenn.edu/LDC2007T03

[3] 誠如評審學者指出，研究資料中 2004 年至 2017 年缺乏標題資料。當時建置資料庫時，這段期間僅蒐集新聞內文並無蒐集標題，十分可惜。我們手動搜尋這段期間的新聞標題，發現了許多符合預期的例子，節錄如下：

(i) a. 互相 餵飯 ， 梅鐸 證實 ， 鄧文迪 偷情 布萊爾 。
b. 無 預警 異動 引 反彈 ， 張惠怡 等 人 嗆聲 中央 ， 將 在 各 鄉鎮 設 連署站。
c. 復仇 南韓 ， 全民 集氣 。
d. 打臉 小英 ， 綠票倉 翻盤 ， 重傷 民主 。

未來在資源許可的情況下將嘗試補齊這段期間的標題資料，以利後續研究。

(17) 資料描述
   a. 新聞資料__標題：1991 年至 2004 年，181 萬 6 千多則（資料一）
   b. 新聞資料__內文一：1991 年至 2004 年，181 萬 6 千多則 （資料二）
   c. 新聞資料__內文二:2004 年至 2017 年,177 萬多則（資料三）

2.2 資料處理與句法剖析

我們使用史丹佛大學中文句法剖析器[4]的剖析結果作為抽取句法結構的重要標記，為最佳化剖析結果，我們進行以下三個資料前處理的步驟：

(18) 資料前處理步驟：
   a. 為標題資料補上標點符號。
   b. 將資料重新斷句。
   c. 使用中研院詞庫小組自然語言處理系統（CKIP CoreNLP）進行斷詞與詞性剖析。[5]

史丹佛大學中文句法剖析器傾向將一個句子的最後一個成分剖析成標點符號，若句子未完整地以標點符號結尾（如標題資料），最後一個詞將被剖析為標點符號，導致剖析錯誤。為解決這個問題，我們將資料一（標題資料）中每筆資料結尾處補上句點符號，如表 1。

[4] https://github.com/UniversalDependencies/UD_Chinese-GSD
[5] http://ckip.iis.sinica.edu.tw/service/corenlp/

表 1：添加標點符號

處理前	處理後
環保署推出環保建設六年計畫	環保署推出環保建設六年計畫。
處理前	處理後
法國不在意日本對其核試的杯葛 豐原醫院貼心手推車服務解決住出院行李煩惱	法國不在意日本對其核試的杯葛。 豐原醫院貼心手推車服務解決住出院行李煩惱。

內文部分（資料二、資料三），中文十億字資料集斷句方式較寬鬆，逗點也是其斷句界標之一，因此資料集中包含了部分不完整句。如表二中左邊欄位的句子，「戈巴契夫曾表示」仍須加接逗點後方的句賓方為完整句，而「愛滋病患者對他社交圈內別人對他的看法」僅包含主語「愛滋病患者」與介賓，卻被視為完整句。為避免句子不完整而造成依存關係剖析錯誤，我們以句點、問號、分號、驚嘆號將資料重新斷句（如表 2）。

表 2：將中文十億字資料集重新斷句

內文	中文十億字資料集斷句	重新斷句
戈巴契夫曾表示，他相信國家安全委員會前首腦克留契柯夫是政變主謀。	1.戈巴契夫曾表示， 2.他相信國家安全委員會前首腦克留契柯夫是政變主謀。	1. 戈巴契夫曾表示，他相信國家安全委員會前首腦克留契柯夫是政變主謀。

內文	中文十億字資料集斷句	重新斷句
愛滋病患者對他社交圈內別人對他的看法，與對待他的方式，非常在意。	1.愛滋病患者對他社交圈內別人對他的看法， 2.與對待他的方式， 3.非常在意。	1. 愛滋病患者對他社交圈內別人對他的看法，與對待他的方式，非常在意。

我們人工對比後，發現中研院詞庫小組自然語言處理系統的斷詞結果優於史丹佛大學中文句法剖析器的斷詞結果，因此，我們先使用中研院詞庫小組 CKIP CoreNLP 系統斷詞與做詞性剖析，再將斷詞結果使用史丹佛大學中文句法剖析器（模型：UD_Chinese-GSD）做依存關係剖析。前處理後的語料庫數據呈現如(19)：

(19) 資料描述（前處理後）
- a. 新聞資料__標題（1991 年至 2004 年）：181 萬 6 千多則，181 萬 6 千多句，1650 萬 2 千多詞 （資料一）
- b. 新聞資料__內文一（1991 年至 2004 年）：181 萬 6 千多則，1699 萬 8 千多句，4 億 6293 萬多個詞 （資料二）
- c. 新聞資料__內文二（2004 年至 2017 年）：177 萬多則，1665 萬 5 千多句，5 億 2 千多萬詞 （資料三）

2.3 抽取依存關係的規則

近年來，很多自然處理的研究使用依存句法分析樹（Universal Dependencies Treebank）來訓練句法剖析器。依存關係（Universal Dependencies） （Nivre et al., 2015, 2016）是

一個跨語言語法的標註框架，相關研究學者訂定了跨語言標註分析的準則（Nivre et al., 2015, 2016），涵蓋的語言超過 70 種，漢語也包含其中。

在依存句法分析的架構下，直接標註詞與詞之間的依存關係，結構較為扁平[6]，如圖 1 所示。

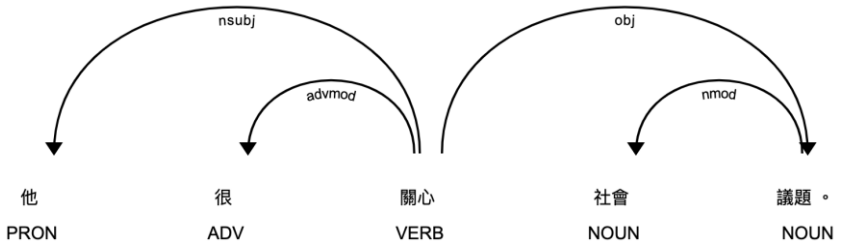

圖 1：依存句法分析

一個句子有一個核心成分（Root），通常為主要動詞，核心成分不依附於（depend on）其他詞，圖 1 的「關心」即為此例，標示的依存關係為“root”，其餘詞則須依附在其他的詞上，以箭頭表示，如「他」、「很」與「議題」皆依附於「關心」，依存關係分別為主詞（nsubj）、副詞修飾語（advmod）以及受詞（obj），而「社會」則依附於「議題」，依存關係為名詞修飾語（nmod）。依存句法樹直接標註詞與詞的關係，而非詞與詞組之間的關係，如圖 1 中依附在「關心」的詞為「議題」，

[6] 詞組結構樹（Constituency tree）是另一個主流的句法分析方式（如：Penn Tree Bank, the Chinese Treebank），樹庫標註詞或詞組結構上的特性（如：NP, VP）與功能上的特性（如：主語 ‘-SBJ’, 時間詞 ‘-TMP’）。詞與詞之間、詞與詞組之間、詞組與詞組之間有不同的語法關係（grammatical relations），有些語法關係透過詞或詞組的組合方式表達（如：Complementation, Adjunction），有些則透過上述功能標注表達（如：predication, modification），整體結構較為立體，較不利於資料的抽取與相關的計算。

而非「社會議題」，如此一來，可直接得知與考察目標詞彙最直接相關的詞有哪些，以及依存關係為何，方便語法的抽取與搭配詞的計算。

本研究目標為考察動詞及物與不及物用法在不同語料庫的對比，需分別抽取以下兩個結構：(1) 動詞 受格論元; (2) 介系詞 旁格論元 動詞。我們依照表 3、表 4 的規則抽取依存關係，組成標的結構，留下例句並統計次數。

表 3：「動詞 受格論元」抽取規則

抽取規則	1. 受格論元依存於目標動詞，兩者依存關係為 obj。 2. 目標動詞與其他詞的依存關係不包含依存關係 mark。 3. 語序限制：受格論元在目標動詞之後。
抽取範例	1. 他 很 關心 社會 議題 。（符合抽取規則，參見圖 1） 抽取詞彙：關心（root） 議題（obj）
濾除範例	1. 他 住 在 台北 。（不符合抽取規則 2，參見圖 2） 住（root） 在（mark） 台北（obj）

表 4：「介系詞 旁格論元 動詞」抽取規則

抽取規則	1. 介系詞 A. 句子包含下欄的介系詞。 B. 依附於目標動詞或是動詞的依附詞。 C. 依附關係為 case 或是 acl。 2. 旁格論元 A. 依附於目標動詞或是介系詞 B. 依附關係為 obj、obl 或是 nmod
介系詞列表	在，以，與，對，從，為，由，到，向，於，和，用，跟，給，至，經，往，靠，替，藉，朝，按，憑，沿，把，對於，幫
抽取範例	1. 我 對 學生 很 關心 。（符合抽取規則，參見圖 3） 抽取詞彙：對（case） 學生（nmod） 關心（root） 2. 我 在 外交界 服務。（符合抽取規則，參見圖 4） 抽取詞彙：在（acl） 外交界（obj） 服務（root）

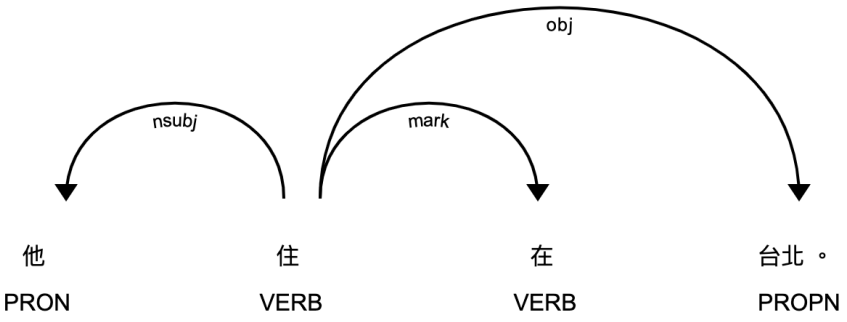

圖 2：依存句法分析

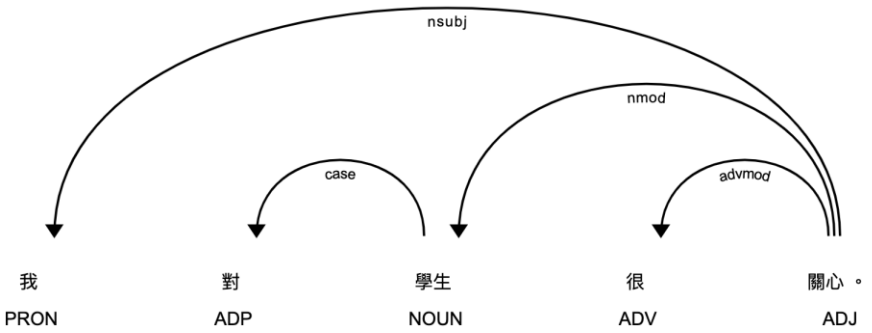

圖 3：依存句法分析

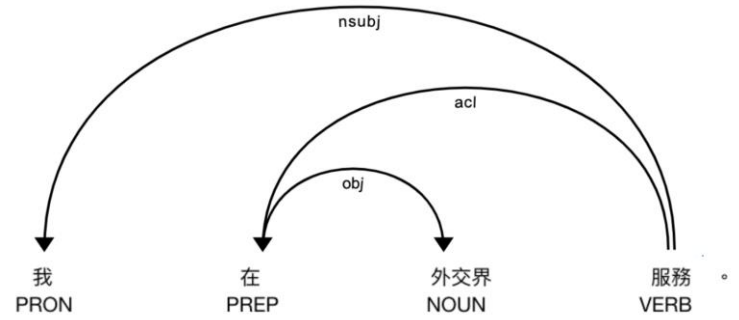

圖 4：依存句法分析

2.4 篩選抽取結果

本研究目標為探討新興的及物化趨勢，聚焦於較非典型的不及物結構及物化現象，因此我們對上節抽取出的不及物與及物結構以及相關例句作更一步篩選，以挑選出適合的例句與統計數據。

首先，我們人工羅列目標動詞與要觀察的介系詞列表，呈現如表 5，剔除未包含這些動詞-介系詞配搭的例句，如(20)。

表 5：動詞-介系詞配搭列表

動詞	搭配介系詞	動詞	搭配介系詞	動詞	搭配介系詞
在意	對, 對於	擔心	對，對於，為, 替	背信	對
關心	對, 對於	斷交	與, 跟, 和	刻薄	對
服務	對，向，為，在, 到	牽手	與, 跟, 和	看好	對, 對於
相見	在, 於	求婚	向，對，跟，和	緊張	對，對於，為, 替
高興	對，對於，為, 替	索賠	向，對，跟	混	到，在，與，和, 跟
劈腿	跟, 與	求救	向，對，跟	前進	向，往，朝

(20) 排除例句

a. 議員雲天寶也到會關心原住民的權益 。

b. 許茹芸本月將赴北極拍 MV，J 以幽默的口吻關心她。

c. 雖然退休了，每天還是從媒體關心雄中。

此外，還需檢查及物結構與不及物結構是否有轉換關係（alternation）。(21a)及物結構「求婚朱海君」可以使用不及物結構「向朱海君求婚」如(22a)，因此(21a)可視為「求婚」及物化結構的實例，但(22b)中的及物結構「索賠九千多萬」無法轉換為不及物結構，意即賓語「九千多萬」無法以動前介賓的方式表達，所以(22b)不能視為「索賠」的及物化結構實例。

(21) a. 之前 Nono 求婚朱海君也曾遭女方媽媽反對。
b. 中油頭屋鄉油管破裂，居民索賠九千多萬。

(22) a. 之前 Nono 向朱海君求婚 ...
b. *居民{向/為/對/…}九千多萬索賠。

為儘可能排除非及物化而來的及物結構，如(21b)，我們分動詞對比賓語與介賓，計算出曾作為賓語與介賓的名詞詞表，例句中的賓語或介賓在名詞詞表中，則留下做後續次數計算，否則則從資料中剔除。

(23) a. 金門縣組成訪問團前進台灣，聯絡鄉誼促銷觀光。
b. 賀伯颱風下午成為強烈颱風向台灣前進。

如(23)中，「台灣」可作為「前進」的賓語與介賓，「台灣」則列為「前進」的搭配名詞詞表之一，因此以「台灣」為論元的例句(23a,b)可作為計算「前進」及物性的資料。

3. 統計結果與分析

我們使用上節的資料處理方法，篩選出目標動詞在三個資料庫中及物結構與不及物結構的例句，為評估系統剖析及物結構與不及物結構的正確度，我們隨機抽取 200 個句子作人工評估，兩位評估者皆受過良好語言學訓練，評估要點為：(1)系統對例句屬於及物結構與不及物結構的判斷是否正確；(2)可否有

及物結構與不及物結構的轉換：若例句屬及物結構，是否可轉換為不及物結構？若屬不及物結構，可否轉換為及物結構。例句符合以上兩項要點才視為正確，否則評估為錯誤，評估結果系統正確率達 77.5%。

我們分資料庫作及物化程度的計算，計算方式為及物結構的句數除以及物結構句數與不及物結構句數的總和。我們使用計算出的結果考察以下兩項漢語及物化的趨勢：

(24) 漢語及物化趨勢
a. 標題比其他書面文體更傾向使用及物句構。
b. 年代較近的資料比較久遠的資料更傾向使用及物句構。

我們以表 5 中的 18 個動詞作為起點，來觀察漢語及物化的趨勢：「在意」、「關心」、「服務」、「相見」、「高興」、「劈腿」、「看好」、「擔心」、「斷交」、「牽手」、「求婚」、「索賠」、「求救」、「背信」、「刻薄」、「緊張」、「混」、「前進」。當中 5 個動詞「在意」、「關心」、「服務」、「擔心」、「看好」的及物化使用現今較普遍、成熟，而資料上也反應出標題較內文的及物性高（如表 6）、年代較近的資料及物性較高（如表 7）的趨勢。

表 6：資料一與資料二及物性統計結果（標題 vs.內文）

	資料一（標題 1991-2004）			資料二（內文 1991-2004）			預測
	及物	不及物	及物性	及物	不及物	及物性	1 > 2
在意	15	1	0.94	436	162	0.73	✓
關心	440	10	0.98	11701	1289	0.90	✓
擔心	237	6	0.98	5255	451	0.92	✓
看好	279	5	0.98	3416	317	0.92	✓
服務	171	30	0.85	6873	4961	0.58	✓

表 7：資料二與資料三及物性統計結果（年代）

	資料二（內文 1991-2004）			資料三（內文 2004-2017）			預測
	及物	不及物	及物性	及物	不及物	及物性	3>2
在意	436	162	0.73	1664	241	0.87	✓
關心	11701	1289	0.90	12545	602	0.95	✓
擔心	5255	451	0.92	9778	538	0.95	✓
看好	3416	317	0.92	6793	214	0.97	✓
服務	6873	4961	0.58	5522	3418	0.62	✓

另有五個動詞的及物性仍在發展中：「斷交」、「索賠」、「混」、「求救」、「前進」，相較發展較成熟的動詞，這些動詞於特殊語境中及物用法在語感上會有爭議。此外，從大數據的角度來看，這些動詞的及物性也較低，如「斷交」的及物性僅有 32%（資料一）、13%（資料二）、11%（資料三），及物化的趨勢也較不穩定。請參照表 8、表 9。

表 8：資料一與資料二及物性統計結果（標題 vs.內文）

	資料一（標題 1991-2004）			資料二（內文 1991-2004）			預測
	及物	不及物	及物性	及物	不及物	及物性	1 > 2
斷交	43	91	0.32	231	1551	0.13	✓
索賠	9	23	0.28	117	369	0.24	✓
混	1	1	0.5	134	82	0.62	✗
求救	11	1	0.92	151	401	0.27	✓
前進	103	2	0.97	1063	477	0.69	✓

表 9：資料二與資料三及物性統計結果（年代）

	資料二（內文 1991-2004）			資料三（內文 2004-2017）			預測
	及物	不及物	及物性	及物	不及物	及物性	3>2
斷交	231	1551	0.13	66	532	0.11	✗
索賠	117	369	0.24	201	261	0.44	✓
混	134	82	0.62	776	371	0.68	✓
求救	151	401	0.27	43	1524	0.03	✗
前進	1063	477	0.69	3011	842	0.78	✓

剩餘的八個動詞及物化的趨勢不清楚：「相見」、「高興」、「劈腿」、「牽手」、「求婚」、「背信」、「刻薄」、「緊張」，但在資料中仍有些有趣的例子，列舉如(25)。

(25) a. 越洋**相見家扶**：頭一遭被認養的兒童飄洋過海和認養者見面 ...（資料三）
b. 他臉書 PO 文**求婚丁文琪**成功 ...（資料三）
c. 我們的一切消費開支很低，但我們不覺得**刻薄**了**自己** ...（資料三）
d. 雖然活動**牽手**了部分企事業單位的**工會、團委**，但此次報名參加的單身男女青年，主要還是以社會報名為主 ...（資料三）
e. 若全民託付黨也支持哈比比，它將會因**背信選民**而自毀前途 ...（資料一）
f. 我知道大家都很**緊張我**，但我更喜歡現在這種狀態的生活 ...（資料三）
g. 很多司法官反彈，但律師公會很**高興**這樣的修法**方向** ...（資料三）
h. 王姓女生雖**劈腿三男**，但一直未遭拆穿。 ...（資料三）

上述的例子證明，這些動詞開始有及物化的現象，只是目前在口語中較普及，尚需時間才能漸漸擴展至書面語資料。

4. 結論與未來展望

前述觀察與研究顯示語法研究應該和資料科學充份結合，取得實證面上的數據支持，並在語言教學和人工智能等面向上發揮其潛在應用價值。事實上，若從中英文平行語料庫做初步觀察，及物化的用法很難找到跨語言的對應，但卻凸顯了中英文在語言類型上的差異：以(26)中的對譯為例，首先是語序上英語的介詞結構在動後而非動前，這與其「中心在前」（head-initial）的特色一致。其次是英語名動互轉的機制非常

發達（亦即 quality ⇒ qualify），而漢語只能靠引介輕動詞（light verb）如「取得」來翻譯。

(26) Andrew qualified as a teacher in 1995.
安德魯 於 1995 年 取得 教師 資格 。

這點在(27)、(28)的對譯中也得到印證：相當於英語中一個簡單的介詞 *for*，中文翻譯卻需要添加「喝」、「找」等動詞，才能形成動後的及物用法：其實此處 *for* 的性質非常接近漢語中語意泛化的輕動詞；一旦翻成中文便需要照顧到賓語的選擇限制，必須用更明確的動詞才行：

(27) It was freezing outside and Marcia longed for a hot drink
外面 很 冷 ， 瑪西雅 很 想 喝 一 杯 熱飲 。

(28) She groped for her glasses on the bedside table
她 在 床頭櫃 上 摸索 著 找 眼鏡 。

此外，此處 *qualify* 其實屬非賓格用法（unaccusative construal），相當於 *to become qualified as a teacher*。由於漢語沒有類似用法，也不太能說「取得資格成教師」，因此「教師」便從補語轉化為賓語名詞前的定語。這些現象若有足夠多的語料可供大數據研究，相信一定能在華語教學、文法寫作、機器翻譯等面向上開啟更具突破性的發展，讓文法理論和應用做更完美的結合。

以上述理念為基礎，我們可以開始發展「客製化」的中文文法搜尋引擎，用大數據理念來研究各種不同文體的語法差異和通則，進而應用到修辭學、文體學、高級寫作、辭典編纂以至符號學、社會語言學等做具有前瞻性的跨界研究。一旦我們有了夠大的資料庫，便可將觸角延伸至認知結構、本體論

（ontology）、歷史語法、語言類型學及普遍語法的研究，讓不同領域的學者既能各取所需又可相互支援。

引用文獻

馮勝利. 2005.〈輕動詞移位與古今漢語的動賓關係〉，《語言科學》4.1: 3-16。

何萬順、蔡維天、張榮興、徐嘉慧、魏美瑤、何德華. 2016.《語言癌不癌？語言學家的看法》。台北：聯經出版社。

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. 2003.〈迎接一個考證學和語言學結合的漢語語法史研究新局面〉，何大安主編《古今通塞：漢語的歷史與發展》，23-47。台北：中央研究院語言學研究所。

蔡維天. 2017.〈及物化、施用結構與輕動詞分析〉，《現代中國語研究》19: 1-13。

齊滬揚. 2000.《現代漢語短語》。上海：華東師範大學出版社。

Chang, P.-C., Tseng, H., Jurafsky, D., and Manning, C. D. 2009. Discriminative reordering with Chinese grammatical relations features. In *Third Workshop on Syntax and Structure in Statistical Translation*, pp. 51–59.

Elming, J., Johannsen, A., Klerke, S., Lapponi, E., Alonso, H. M., & Søgaard, A. 2013. Down-stream effects of tree-to-dependency conversions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 617-626.

de Marneffe, Marie-Catherine, and Christopher D. Manning. 2008. *Stanford typed dependencies manual*. Technical report, Stanford University.

de Marneffe, M.-C., Connor, M., Silveira, N., Bowman, S. R., Dozat, T., and Manning, C. D. 2013. More constructions, more genres: Extending Stanford dependencies. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pp. 187-196.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC), pp. 4585-4592.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Ta¨ckstro¨m, O., Bedini, C., Bertomeu Castello´, N., and Lee, J. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL), volume 2: Short Papers, pp. 92-97.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajicˇ, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. 2016. Universal dependencies v1: A multilingual tree- bank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC), pp. 1659-1666.

[Received 28 February 2020; revised 11 September 2020; accepted 31 October 2020]

蔡維天
國立清華大學
語言學研究所
wttsai@mx.nthu.edu.tw

楊謦瑜
國立中興大學
外國語文學系
chingyu@dragon.nchu.edu.tw

陳映竹
國立清華大學
資工系
jocelyn@nlplab.cc

陳志杰
國立清華大學
資工系
jjc@nlplab.cc

張俊盛
國立清華大學
資工系
jason@nlplab.cc

A DATA SCIENTIFIC STUDY OF TRANSITIVIZATION IN CHINESE

Wei-Tien Dylan Tsai[1], Ching-Yu Helen Yang[2],
Chen Ying-Zhu[1], Jhih-Jie Chen[1], and Jason S. Chang[1]
[1]*National Tsing Hua University*
[2]*National Chung Hsing University*

ABSTRACT

From the perspective of data science, this study aims to investigate an emerging phenomenon called "transitivization" in Mandarin Chinese. It is a syntactic change that turns a preverbal applicative argument into a postverbal direct object. This process can be viewed as a "renaissance" wei dong shi 'beneficiary verb form' in Classical Chinese. It may either die out after a short period of time, or have an explosive growth in its popularity just like the so-called "language cancer" recently observed in Taiwan. Therefore, we need to address the issues by combining data science and syntactic analyses. This move enables us not only to give a comprehensive review of the current status of this syntactic change, but also to make a plausible prediction about its future development. Finally, it is instructive to note that transitivization is found only in certain stylistic registers like news headlines, which in turn allows us to look deeper into the pragmatic considerations and underlying mechanism of the whole process.

Keywords: transitivization, light verbs, Chinese syntax, data science