The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 424-433

標註英中同步樣式文法之研究

Annotating Synchronous Grammar Patterns across English and Chinese

楊謦瑜　Ching-Yu Helen Yang, 陳映竹　Chen Ying-Zhu, 張俊盛　Jason S. Chang
國立清華大學資工系
Department of Computer Science
National Tsing Hua University
chingyu@nlplab.cc, jocelyn@nlplab.cc, jason@nlplab.cc

林依蒨　Yi-Chien Lin
國立清華大學外語系
Department of Foreign Languages
National Tsing Hua University
nicayclin@gmail.com

蔡維天　Wei-Tien Dylan Tsai
國立清華大學語言學研究所
Linguistics Institute
National Tsing Hua University
wttsai@mx.nthu.edu.tw

摘要

本文從語料庫與機器學習的角度，來進行英語和華語的同步文法研究。我們認為這種對比性研究，可以借助既有的英語樣式文法的研究成果，提供華語辭典學、華語教學新的研發方向。在我們的研究路線上，運用了辭典中的雙語例句，來發掘英語、華語的動詞句法規則。我們的方法涉及自動辨識例句中的英語文法規則、運用詞彙對應的技巧產生華語規則的建議，最後，透過人為分析產生正確英華語法規則的資料集。我們把這個研究方法，運用在劍橋大學出版社的線上英漢辭典的例句，初步完成英語「動介賓」規則的華語對應規則的分析。我們就初步的研究結果，說明標註華語文法規則的指導原則，觀察分析所得到的華語文法規則的統計分布。最後完成的資料集，可望有助於提供華語文法規則自動擷取的機器學習研究。

關鍵詞: 樣式文法 pattern grammar，同步文法 synchronous grammar，自動文法推導 grammar induction

一、緒論

實證性、語料庫為本的句法研究，有幾個不同的作法。最常見的方式，透過抽樣取少部分句子樣本，人為分析這些句子的句法結構，以建構剖析樹庫的方式，來得到一個文法剖析的代表性樣本（Marcus, Santorini, and Marcinkiewicz 1993）。另外一條詞彙式的研究路線（Sinclair 2000），是分析語料庫中個別詞彙的樣本，分析其常見文法規則，以羅列詞彙化文法規則。在華語句法研究上，已經有中研院的樹庫，Chinese Treebank 8.0 等樹庫資料集的研究發展（Huang, et al. 2000; Xue and Palmer 2003, Xue, et al. 2005）。然而，華語詞彙化文法（如樣式文法 pattern grammar），卻相當缺乏相關資料與研究。

本文描述一項計畫，透過英語既有的詞彙化文法規則，以及相關的平行雙語句子，標示華語句子中出現的對應規則。其目的在於產生對應的華語文法規則的小量訓練資料，以利後續可以饋入機器學習系統，在單語語料庫中，擷取更全面的華語文法規則。

樣式文法（Pattern Grammar ）是一種描述個別詞彙的句法環境的語言模型。PG 源自在大型語料庫中，觀察個別詞彙的實例。在知名的考林斯出版社與伯明罕大學的 COBUILD 計畫中，Hunston, Francis, and Manning (1996, 1997) 發展出的文法理論，和一般習知的片語結構文法（phrase structure context-free grammar），有很大的區隔[1]。

PG 主張，語言學家、辭典學家可以針對每個實詞（動詞、名詞、形容詞），觀察語料庫來賦與一組文法規則（patterns），用以描述該詞彙的主要用法。PG 更進一步主張，通常規則有類似語意。PG 的符號形式，和 CFG 有些不同：

- 文法規則由一串符號構成，其中有個大寫符號（V, N, 或 ADJ ）代表中心語。其餘的符號為小寫的文法結構（'v', ‘n’, ‘adj’, ‘adv’, ‘prep’ 分別代表動詞片語、名詞片語、形容詞片語、副詞片語、介詞），但是也可以是特定的介詞。這些小寫的元素（個別，或整體而言）可以視為中心詞的句法搭配（grammatical collocations 或 colligations）。

[1] https://www.collinsdictionary.com/word-lovers-blog/new/what-are-grammar-patterns,524,HCB.html

- 小寫元素除了上述的幾種之外，還可以是句法結構（that 子句）、動詞型態（如 to-inf, inf, v-ing, v-ed, passive）、虛詞（wh, ord）語意搭配（如 amount, color, number, name）或者表面詞彙（如 get, be, way）。

所以，Pattern Grammar 是一種線性編碼，用以把通常的多層 CFG 規則的結構壓扁成為一層，並強調實詞（如動詞）與虛詞、文法結構（如副詞、子句）搭配的現象，可以直覺地溝通字詞的用法，明確地說明文法現象。因此，很適合作為辭典、教學之用。PG 最早應用在 Collins COBUILD English Dictionary (1995) 來標示詞條下每個詞意的文法編碼（Grammar coding）提供一種簡單容易理解的符號形式。然而，雖然簡明易解，PG 又是非常有彈性，有豐富的表達能力。而華語詞彙知識庫(如中研院 eHowNet)有句法訊息，但是並沒有透過嚴密的語料庫語言學分析，所以其文法規則的涵蓋不完備。有鑑於此，我們認為有必要透過英語文法規則的投射(projection) 來加速華語詞彙文法的研究，以提升未來華語辭典的文法方面的教學效果。

二、相關研究

用辭典、文法書、分級讀本來學習語言，有悠久的歷史，也是直覺合理的作法。然而，Sinclair (1991) 指出辭典傳統上過於重視詞意的解釋，而忽略了文法、語用的說明，也缺乏文體、領域的資訊。所以，Sinclair 主張運語料庫語言學，計算辭典學，詞彙索引典，來改善語言學習，和參考工具書的編輯。在知名的 Collins 出版社和 Birmingham 大學的 COBUILD 合作計畫下，Hunston, Francis, and Manning (1996, 1998) 出版了兩冊以詞彙為中心（lexical approach or lexical syllabus）的文法規則彙編——Collins COBUILD Grammar Patterns: Volumes 1 and 2。

在語料庫為本的研究上，Weber (2001) 描述如何運用語料庫和學生習作的交互作用，來幫助法學院學生的法律寫作。Sun (2007) 分析並標示語料庫，開發 *Scholarly Writing Template* (SWT) 系統，來幫助研究生寫作。我們聚焦在動詞的文法規則，不論是否和常見的寫作樣板有關。

在自動寫作評分的研究上，美國的 Education Test Service (ETS) 用機器學習與統計方法開發了 *Criterion* 系統 (Burstein, Chodorow and Leacock, 2003) 可以提供包含文法錯誤的寫作回饋。該系統已經用於對 4 to 12 年級的學生寫作，以及 TOFEL 和 GRE 的作文測驗部分。我們的研究集中在英語、華語的動詞文法規則，可以用來檢驗語言學習者最容易觸犯的動詞錯誤。

Chang and Chang (2015) 提出一個輕度督導式方法，自動推導英文的文法規則，並透夠提示來輔助學生寫作。這個系統的構想，延伸搜尋（如 *Google Suggest*）、翻譯（如 *TransType*）中的自動提示與完成的功能(Langlais, Foster and Lapalme, 2002)。

有別於先前的研究，我們提出一套開發文法標註資料集的作法，可以輔助語言分析師，標註資料集，以提供英語、華語文法，甚至雙語同步文法，自動推導的訓練資料，以期有助於華語，以及雙語同步文法的語言學，語言工程研究。

三、建構英華同步文法標註資料集

為了推導出華語文法規則，我們打算利用英華雙語例句，參照既有的英語文法規則，斯尋文法規則出現的實例，選取相關例句，並在例句的華語部分標註對應的華語文法規則。這項工作，並不如想像中那麼簡單。平行語料庫中的句子，常常結構過於複雜，有時也不容易找到特定既有英語文法規則的實例。所以，我們採取雙語辭典中的代表性例句，逕行此項工作。我們的方法，有三個步驟：產生雙語文法規則標示的草稿、人工標註、分析標註的結果。以下分別敘述之。

3.1 產生雙語標示詞料的草稿

我們將介紹一個將英語的動詞文法規則（grammar pattern） 找出對應的華語文法規則的方法。文法規則的資料來源為 Collins COBUILD Grammar Patterns: Verbs（arts-ccr-002.bham.ac.uk/ccr/patgram/）書中第二章節所列的規則，共 34 條。我們擷取符合 34 條規則的動詞。接著，我們蒐集劍橋大學線上英漢字典的例句（dictionary.cambridge.org/us/dictionary/english-chinese-traditional/）。我們先透過預處理辭例句，選出含相關動詞（如 talk）以及相關文法規則（我們透過相依關係分析，確認有 **V about n** 的句法結構，並擷取英文介詞、賓語）。接著，我們運用詞彙對應（word

alignment）的技巧，產生動詞、介詞、賓語的華語對應詞。憑藉著這些資料，我們產生如下格式的資料。標記的資料格式如下：資料點標號、四項資訊標號，四項資訊（包括英語句、華語句、英語實例＋文法規則、華語實例＋文法規則。以下顯現一筆 TALK: V about n 的例子：

1.1 ||| We were just talking about Gareth's new girlfriend

1.2 ||| 我們 剛才 在 談論 葛瑞 的 新 女友 。

1.3 ||| talk about girlfriend ||| talk : V about n

1.4 ||| 談論 女友 ||| 談論 : V n

第三項的動詞以及規則也為已知的資料，但仍須人工再確認是否和句子結構有相符。

第三項的英語詞組抽取方式為，先將 英語句子使用 spaCy 標記詞性，再抓取動詞與介系詞後面的名詞。其餘的項目為需要人工標記的部分，皆先使用 heuristic 的方式得到一個暫時的結果，再由人工校正。

第四項的華語動詞透過事先訓練好的中英雙語 word2vec model，計算英語動詞與華語句中詞語的相似度，找出相似度最高的華語，作為英語動詞的翻譯。華語詞組抽取即是透過找出的華語動詞，經由 Stanford Parser 得到華語句子的相依關係（dependency），找到與動詞相依的介系詞和名詞。

3.2 資料集標註指南

人工標記校正的準則，英文、中文分別列舉如下：

3.2.1 英語標註原則

應正確地標示出規則的實例，確認「動詞、介詞、賓語」是否正確，並確認由介詞與賓語組成的介賓詞組為動詞的必備成分。以下按賓語的詞類分項說明標注原則：

(a) 賓語若為名詞組：確認抽取實例是否為名詞組的中心語，而非修飾語或是所有格。如例子 (1)。

(1) We were just **talking about** Gareth's new **girlfriend**.

(b) 賓語若為動詞組，實例應擷取動詞組中的動詞原型。如例子 (2)。

(2) I'm **thinking about buying** a new car.

(c) 賓語若為疑問子句，應該標示 *wh*-詞以及 “to” (若句子含有 “to”)。如例子(3)。

(3) We couldn't **agree on what to buy**.

(d) 賓語若多於 1 個字, 則用底線連接，如例句 (4)。

(4) Who is going to **speak for** (= represent in a court of law) **the accused**.

此外，依照文法規則實例的情況，標示文法規則的元素，包含中心詞、V、介詞、賓語類型，如例句 (5) – (8)。

(5) We were just talking about Gareth's new girlfriend. （talk : V about n）

(6) I'm thinking about buying a new car. (think : V about ving)

(7) She was dithering about what to wear. (dither:V about wh_to-inf)

(8) Nick was enthusing about how well things worked out. (enthuse:V about wh)

3.2.2 華語標註原則

在華語詞組實例抽取的部分，與英文一樣，主要任務為確認「介系詞」、「賓語」以及「動詞」，並確認由介詞與賓語組成的介賓詞組為動詞的必備成分，標記資料中部分資料的斷詞不理想，但考慮後續訓練模型方便，標記時盡量維持原始資料的斷詞，不做修改。中文與英文的詞彙語法不同，有中英文能直接對應的語料，但有許多語料無法直接對應，因此華語標示原則在力求保持中英語料的平行性，以求能區分華語不同結構。以下分賓語、動詞進行討論。

3.2.3 華語文法規則標註原則

依照文法規則實例的情況，標示文法規則的元素，包含中心詞、V、介詞、賓語類型，中心詞以動詞實例標注，介系詞以介系詞實例標注，如(27-29)。

(27) 下 一 位 講 者 將 **就** 瀕危 **昆蟲 作 報告** 。（作_報告:就 n V）

(28) Andrew **qualified as** a **teacher** in 1995.

安德魯 於 1995 年 **取得 教師 資格** 。 (取得^資格 : V n)

(29) 餐館 裏 所有 的 人 都 **擁 到 他們 周圍** 唱 了 起來 。(擁:V 到^周圍 n)

若英文介系詞對應到華語動詞，則標註為 “v n”，如(30)呈現。

(30) Janet is **speaking for** the **motion**.
珍妮 特 **發言 支持** 這 項 **動議**。 (發言 : V v n)

賓語則按詞類標註：

(a) 賓語為名詞，標註為 n，如(27-30)。

(b) 賓語為動詞組，若英語有相對應的動詞，標示為 vp 如(31)，否則標註為 v_n 如(32)。

(31) I'm **thinking about buying** a new car.
我 在 **考慮 買** 輛 新 車 。 (考慮 : V vp)

(32) She **groped for** her **glasses** on the bedside table.
她 在 床頭櫃 上 **摸索** 著 **找 眼鏡** 。(摸索 : V v_n)

(a) 賓語為疑問子句時，若疑問詞為子句的謂語，直接標註「疑問詞」，如(33)，若不為子句的謂語，則標註為「疑問詞_v」，如(34)。

(33) 你 **覺得** 這 個 改善 地鐵 系統 的 最 新 **計劃 怎麼樣** ？(覺得:V 怎麼樣)

(34) 你 就 **假裝** 好像 **什麼** 事 都 沒 **發生** 過 。 (假裝 : V 什麼_v)

(d) 英語賓語對應到華語賓語的修飾語，如(35)，應標註為 “n_的_n”。

(35) Did you **ask about** the **money**?
你 **打聽 錢 的 事** 了 嗎 ？ （打聽 : V n_的_n）

四、初步標示結果的摘要分析

英語動介賓的文法規則（包含各種介詞）對應到的華語文法規則，有以下幾個現象：

1. 視介詞的不同（如 about 和 against），其對應的華語規則，有很大差異。所以，文法規則凸顯個別介詞，是很有必要的。對機器學習、語言學習都會有比較好的效果。

2. 一般而言，不論介詞為何，英語不及物動詞的 V p n 的文法規則，對應到華語動詞

文法，集中在不及物轉為及物（V n），介詞片語往前移動（p n V），或規則不變（V p n）三大類。而英語介詞（p）所對應的華語介詞，變化範圍也不大。

3. 少部分的介詞（如 against, for）有很強的傾向，對應到華語動詞（如「反對」、「支持」、「贊成」等）。如此一來，對應的華語規則，就變成 V v n。

4. 對應到的華語介詞有有一些現象和英語介詞不同。具體而言，英語介詞片語，對應到華語常常傾向於「介詞＋名詞＋方位詞」的型態。例如 CHASE: V after n 對應到「追：在 n 後面 V」。為了表示「在」和「後面」的成雙成對的關係，我們改寫為「在^後面 n V」。其中的「^」符號，表示其後的 「n」插入「在」和「後面」兩者之間。類似辭典常用的 「在～後面」或「在 … 後面」的表達方式。

5. 歸納起來，有近 6 成，對應的華語文法規則，是 p n V 或 V p n。另外，有三成多為 V v n。介詞以「在」或「在^方位詞」為多，其餘的介詞包括「向、為、冲、朝、對、就、與、跟、和、到鬧、於、給、替」等等。

五、結論

未來有許多方向可以繼續探索，並改進目前的作法和結果。例如，目前抽取英文規則的方法還可以考慮用搭配的統計分析，篩選教具代表性的例子，同時規避剖析錯誤，抽取了不正確的英文規則。我們也可以透過更大量的雙語資料分析，得到比較正確的華語文法規則的建議，減低語言分析師的工作負荷。另外一個有趣的研究方向，是擴大英語、華語文法規則的範圍，以包括更複雜的句法現象。更進一步的研究，應該從動詞，延伸到名詞、形容詞。還有一個更重要的研究方向，是用完成的資料集，訓練一個自動產生雙語同步文法規則的系統。該系統的結果，可以作為編撰華語辭典，開發機器翻譯系統的基礎。

總結起來，我們呈現一個方法，可以採電腦輔助的方式，開發同步文法規則資料集。我們的研究路線涉及運用了辭典中的雙語例句, 來發掘英語、華語的動詞的對應句法規則。我們的方法有三個步驟：自動辨識例句中的英語文法規則、運用詞彙對應的技巧產生華語規則的建議、人為分析產生正確英華同步文法規則的資料集。我們把這個研究方法, 運用在劍橋大學出版社的線上英漢辭典的例句, 初步完成英語「動介賓」規則的華

語對應規則的分析。我們就初步的研究結果，說明標註華語文法規則的指導原則，觀察分析所得到的華語文法規則的統計分布。最後完成的資料集，可望有助於提供華語文法規則自動擷取的機器學習研究。

參考文獻

[1] Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank." (1993).

[2] Sinclair, John. "Lexical grammar." *Naujoji Metodologija* 24 (2000): 191-203.

[3] Hunston, Susan, and Gill Francis. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Vol. 4. John Benjamins Publishing, 2000.

[4] Huang, Chu-Ren, et al. "Sinica Treebank: design criteria, annotation guidelines, and on-line interface." *Second Chinese Language Processing Workshop*. 2000.

[5] Xue, Naiwen, et al. "The Penn Chinese TreeBank: Phrase structure annotation of a large corpus." *Natural language engineering* 11.2 (2005): 207-238.

[6] Xue, Nianwen, and Martha Palmer. "Annotating the propositions in the Penn Chinese Treebank." *Proc of the 2nd SIGHAN workshop on Chinese language processing*, 2003.

[7] Johnson, Christopher (2000). "Review of Pattern grammar: A corpus-driven approach to the lexical grammar of English". Computational Linguistics. **27** (4): 318–320.

[8] Francis, Gill; Hunston, Susan; Manning, Elizabeth, Collins COBUILD Grammar Patterns 1: Verbs, HarperCollins, 1996.

[9] Francis, Gill; Hunston, Susan; Manning, Elizabeth, Collins COBUILD Grammar Patterns 2: Nouns and Adjectives, HarperCollins, 1997.

[10] Hunston, Susan; Francis, Gill, Pattern Grammar: A corpus-driven approach to the lexical grammar of English, John Benjamins, 2000.

[11] Francis, G. (1993). A corpus-driven approach to grammar – principles, methods and examples. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds). *Text and Technology: in Honour of John Sinclair*. Amsterdam: Benjamins, pp. 137–156.

Francis, G., Hunston, S. & Manning, E. (1996). *Collins COBUILD Grammar Patterns 1:*

Verbs. London: HarperCollins. Francis, G., Hunston, S. & Manning, E. (1998). *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

[12] Hunston, S. & Francis, G. (2000). Pattern Grammar: A corpus-driven approach to the lexical grammar of English. Amsterdam: Benjamins.