國科會大專生研究計畫 計畫書

Programming Linguistic Competence:

The Case of Chinese Approximative Adverbs “*Chayidian*”, “*Chabuduo*” and “*Jihu*”

**語感判斷的程式化工程：**

**漢語近似副詞「差一點」、「差不多」和「幾乎」**

學生:國立陽明交通大學 外國語文學系 陳畯田

中華民國112年2月13日

# 摘要

「差一點」、「差不多」和「幾乎」三漢語近似副詞為易混淆的近義詞，先前已有許多針對漢語近似副詞，以語料庫為基礎的相關研究 (如Chuang 2017; Shyu 2022)；Chuang (2017) 以統計搭配詞角度切入，而Shyu (2022) 則歸納部分語料反映之使用通則試圖捕捉「差一點」、「差不多」和「幾乎」的語意差異及句法分佈。然而此兩類方法均受語料完備程度、語感穩定性、人工操作實驗的有限效率和涵蓋之語料量等因素影響，存在諸多先天限制，對近義詞差異的解釋侷限於部分語料的現象歸納結果，缺乏具一致性的普遍解釋。本計畫旨在利用以 X-bar Theory (Jackendoff 1974, 1977) 驅動的 NLP/NLU 系統 Articut / Loki (Wang et al. 2019)[[1]](#footnote-1) 探勘「差一點」、「差不多」和「幾乎」三漢語近似副詞句法結構使用限制，透過自動化檢閱分析大量語料提升針對近義詞句法結構使用限制解釋的涵蓋性，實現將語感程式化執行；在提升處理效率和涵蓋語料量的同時，也解決語感不穩定造成的潛在問題，最終生成具備一致性操作流程的語句合法判斷的運算程式模型。

# 研究動機與問題

漢語近似副詞「差一點」、「差不多」和「幾乎」 (如例(1)) 為漢語中易混淆的近義詞 (Chuang 2017; Shyu 2022) ，此三近似副詞在語意上接近英文副詞「almost」(Chuang 2017)，然而各自表現的語意不同，在使用上也須遵守不同結構限制 (如例(2))，易造成非漢語母語者在學習上的困境，先前已有不少以語料庫為基礎之漢語近似副詞相關研究。

1. ) 漢語近似副詞：
   1. 老王 **差一點** 吃了三碗白飯。
   2. 老王 **差不多** 吃了三碗白飯。
   3. 老王 **幾乎** 吃了三碗白飯。
2. ) a. 我和他身高 **差不多**。
   1. 我和他身高 **差一點**。
   2. \*我和他身高 **幾乎**。

現行以語料庫為基礎的近義詞研究 (Chuang 2017; Shyu, 2022)，多以人力進行以下(3)-(5)的操作，並傾向以統計搭配詞或記錄使用情形的方式解釋語言現象。

1. ) 前處理階段：
   1. 從大量語料中抽樣並逐一審查語料。
   2. 清洗語料以排除雜訊。
   3. 區分符合和不符合語法的語料。
2. ) 實驗階段：
   1. 先將語料收斂成句法結構態樣 (Syntactic Pattern) 進行比較。
   2. 依據比較結果進行分析和歸納。
   3. 以文字或表格紀錄判斷依據和操作過程以提出假說。
3. ) 驗證階段：
   1. 從其餘語料中抽樣並逐句依先前判斷依據和操作過程進行驗證
   2. 若假說無法解釋語料所反映之現象則須修正假說。
   3. 反覆驗證直至假說涵蓋語料所反映之現象。

然而，近義詞研究若只以語料庫為基礎，執行研究步驟(3)-(5)會有以下三點先天限制。

1. 語料完備性：

以語料庫為基礎的近義詞研究受限於語料蒐集的完備性。以中央研究院漢語平衡語料庫為例，其僅蒐集1981年至2007年間的語料，包含「差不多」之語料總計 405 筆，而含「差一點」者僅44筆，此兩者數量差異大之外，少量語料所反映之近義詞使用現象涵蓋性亦有限。

1. 語感穩定性：

上述提及之操作步驟多為語感所驅動，然而語感易受個體認知差異及時空差異影響，且語感疲乏 (language fatigue) 導致的不穩定性也影響實驗操作的一致性，進而提升結果驗證的難度。

1. 處理效率：

處理語料、操作實驗和驗證語料的工作若透過人力執行，將無法大量批次處理，以至於處理效率和數量有限，近而難以透過語料庫外的大量語料來提升解釋的涵蓋性或針對語言系統本身的持續變化進行修正。

本計畫旨在以現行語料庫近義詞研究 (Chuang 2017; Shyu, 2022) 為基礎，針對漢語近似副詞「差一點」、「差不多」和「幾乎」之句法分佈以及語意特性的語感判斷設計出明確程式化的程式碼，企圖實現批次自動化驗證大量語料；以大量語料提升解釋涵蓋性的同時，程式碼的穩定運作也得以避免語感不穩定性造成的影響，最終生成以句法結構驅動 (rule-driven)、實現以語言學角度理解漢語近似副詞使用規則的程式模型。

# 文獻回顧與探討

近似副詞在語意上包含兩個語意成分：近似成分 (proximal component) 和極性成分 (polar component) (Chuang 2017; Shyu, 2022)。舉例來說，在(6)之中，「差不多」的近似成份代表Lebron James接近得分王的狀態了，而極性成分則表示說話的時候，尚未達到此狀態，即「Lebron James今天成為得分王」此命題為假。

1. ) Lebron James今天 **差一點** 成為得分王。

漢語近似副詞「差一點」、「差不多」和「幾乎」均具近似成分 (proximal component) 和極性成分 (polar component) (如上節例(1))，然而此三近似副詞各自有不同解讀意義和使用限制。此節目標有二，闡明「差一點」、「差不多」和「幾乎」相異之語意及句法結構特性，並依先前研究歸納此三近似副詞之句法結構分布，評估程式化的可能性。

## 近似成分和極性成分

「差一點」、「差不多」和「幾乎」在語意上均可表達「接近某一狀態，卻非其狀態」 (Shen 1987)，亦即一命題的反面或該命題不為真 (如例(7))。(7abc)均可表當前情況接近「張三完成作業」一命題，但「張三完成作業」一命題不為真。

1. ) a. 張三 **差一點** 完成作業。
   1. 張三 **差不多** 完成作業。
   2. 張三 **幾乎** 完成作業。

與(7)之中的完成動詞搭配時，三種近似副詞都具備近似成分和極性成分，但Shyu (2022) 觀察到與狀態謂語搭配時，「差不多」表達近似的方式具三種可能的解讀。 以(8a)為例，「小明的身高」可能略為少於、多於或等於 185 公分，但「差一點」與「幾乎」都只能表達略為少於185 公分的語意，因此在此語境中，「差不多」可以不具極性語意，即小明的身高可以等於 185 公分。

1. ) a. 小明的身高 **差不多** 有185公分。
   1. 小明的身高 **差一點** 有185公分。
   2. 小明的身高 **幾乎** 有185公分。

此外，不同漢語近似副詞所蘊含之成分程度也不盡相同。相較於「幾乎」，「差一點」的極性成分帶有更多負面意向解讀，即「差一點」較傾向否定一事件的發生 (Chuang 2017)，如例(9)。

1. ) a. 他 **幾乎** 死了，快買棺材吧。
   1. \*他 **差一點** 死了，快買棺材吧。

(9ab)中，相較於「幾乎」，「差一點」較強的極性成分表示當前情況為命題「他死了」的反面，與此一命題相距較遠，造成前句與後句「快買棺材吧」內容之邏輯不相符。而「差一點」與「差不多」側重的語意不同，以(10)為例，「差一點」相較於「差不多」，更強調其極性成分 (即李四還沒到終點)，「差不多」則強調近似成分 (即李四離終點很近)。

1. ) a. 李四 **差一點** 到終點了。
   1. 李四 **差不多** 到終點了。

## 句法結構限制

就句法結構分布而言，「差一點」、「差不多」和「幾乎」各自受到不同限制，在文句中不可完全交替使用，透過觀察語料能認識其三近似副詞的部分句法結構限制與使用情形。

* 主要謂語

首先，「差不多」本身可以作為主要謂語使用，其可能位於比較句句末一名詞組後方，而「差一點」和「幾乎」則不可以 (如例(11))。

1. ) a. 沿海和內陸工資 **差不多**。
   1. \*沿海和內陸工資 **差一點。**
   2. \*沿海和內陸工資 **幾乎**。 (Shyu, 2022)

* 終結語意

「差一點」、「差不多」和「幾乎」不可以被用來修飾不具終結語意 (telicity) 的謂語。

1. ) a. \*這個孩子 **差一點** 高了。
   1. \*這個孩子 **差不多** 高了。
   2. \*這個孩子 **幾乎** 高了。
2. ) a. 衣服 **差一點** 乾了。
   1. 衣服 **差不多** 乾了。
   2. 衣服 **幾乎** 乾了。

在(12abc) 中，「高」本身為 open-scale adjective，沒有最大值及最小值，其不具終結語意，則無法被「差一點」、「差不多」和「幾乎」修飾。反觀(13abc) 中，「乾」本身為帶有終結語意的closed adjective (Chuang 2017)。

此外，closed adjective 亦分為total adjective 及 partial adjective，前者只能表示在程度量表上的最低值，其具有一標準值，「乾」即為 total adjective，其意義為濕度 (humidity) 量表上的最低值；而「濕」則為 partial adjective ，其無標準值，表示意義可以為濕度 (humidity) 量表上的任何一個值。(Chuang 2017)。根據此細微的語意差異，Chuang (2017) 發現「差一點」可以修飾partial adjective，而「差不多」則不可以 (如例(14))。

1. ) a. 衣服 **差一點** 濕了。
   1. \*衣服**差不多** 濕了。

(14a)中，當前衣服狀態正接近濕度量表上的最大值，因「濕」為partial adjective，(14a)中衣服狀態仍可以為濕度量表上的任意值，然而受「差一點」語意特性的影響，衣服狀態則可以被確定為濕度量表上的最小值，反觀(14)b)，以「差不多」修飾 partial adjective「濕」，將會因為「差不多」針對程度和值無特定明確之表示意義，其可能表示濕度量表上的任何值，造成解讀上的困難。

Shyu (2022)近一步觀察到「差不多」和「幾乎」在修飾不具終結語意之謂語時，需要量詞 (quantifier) 給予程度衡量語意上的有限值，以補齊缺少的終結語意 (如例(15)-(16))。

1. ) a. \*他 **幾乎** 玩。
   1. 他 **幾乎** 天天玩。
2. ) a. \*衣服 **差不多** 濕了。
   1. 衣服 **差不多** 全濕了。

以上所提及之句法結構限制，為先前相關研究 (Chuang 2017; Shyu 2022) 透過觀察語料所歸納統整，其中句法結構限制受到「差一點」、「差不多」和「幾乎」本身語意特性影響。其中，一近似副詞是否被當作主要謂語，可以透過分析句法結構態樣 (Syntactic Pattern)，抽取位於句末一名詞組之後的詞組得知，而句中被近似副詞修飾之謂語是否具有終結語意之部分判斷，亦可從一詞組本身構詞結構中得知 (如例(11))，此二重要結構限制均可利用以 X-bar Theory 驅動的 NLP/NLU 系統 Loki/Articut 輔助程式化執行，經過中文斷詞、加註詞性標記和適當的轉換後，語料之於電腦即在保留語言結構的前提之下成為可以操作的單位，再逐步將判斷依據以程式碼撰寫即可將上述限制交由電腦判斷彙整，實現語感的程式化。

# 研究方法與步驟

## 研究方法 – Articut/Loki as rule-driven NLP/NLU programs

理論語言學的理論建構目的在於針對語感提出可解釋的明確步驟和形式化後的判斷過程做為依據，以上特質賦予語感以程式碼撰寫執行 (programmable) 的潛質。此研究將觀察中研院平衡語料庫中相對少量的語料 (含「差一點」、「差不多」、「幾乎」之語料分別為 44 筆、405筆及2185筆)，將自然語言在保有結構本質的前提下，轉化為電腦可運算的單位，再以程式碼撰寫判斷依據和操作流程模擬語感的運作，生成可執行的程式模型，並將Sketch Engine中的大量語料放進模型中進行驗證並針對結果進行修正 (含「差一點」、「差不多」、「幾乎」之語料各10000筆)，將操作實驗的過程批次自動化執行並最大化語料驗證的數量及效率，提升解釋涵蓋性。此外，以程式碼撰寫生成的模型具有一致性，可以避免研究過程中語感受到個體及時空差異影響。

Articut/Loki (Wang et al. 2019) 為根據生成語言學原理建構的 NLP/NLU 系統， 能在保有自然語言結構本質的前提下完成運算。Articut 不同於現行以統計機率計算詞頻或語料庫字典驅動的斷詞系統 (Jieba、CKIP tagger)，其以X-Bar Theory 為基礎運作，在完整保留功能詞 (functional words) 的前提之下，以其為核心語 (functional heads)，將句子套用到不同句法樹進行特徵檢查 (feature checking)，並在執行中文斷詞 (Chinese Word Segment，以下簡稱 CWS) 的同時依節點加註詞性標記 (Part of Speech, 以下簡稱 POS)，其過程中考量到漢語中雙音詞及非核心語重音等構詞上的特性。而 Loki 將經 Articut 標記斷詞的語句以正規表示式 (Regular Expression, 以下稱 Regex) 輸出，以句法及語意結構做為運算操作符號，在不同的語境和意圖之下，比對語料結構，使用者可於相關程式區塊中撰寫程式碼，以針對不同結構進行詞彙參數抽取，設計回應等操作。

## 研究步驟

1. 將語料轉化為電腦可運算的單位：

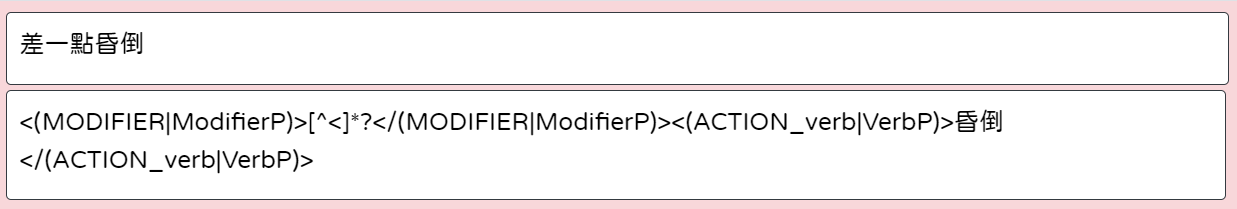
首先，使用Articut將語料進行中文斷詞 (CWS) 和詞性標記 (POS) 處理，而CWS和POS 賦予語料具一致性的操作符號，範例操作如(17)：

1. ) Articut的CWS與POS操作範例：



接著利用Loki Advanced 將以上結果輸出為以Regex表示之句法結構態樣 (Syntactic Pattern)，此步驟將(17)斷的詞標記結果轉為電腦可運算的單位。。範例操作如(18)：

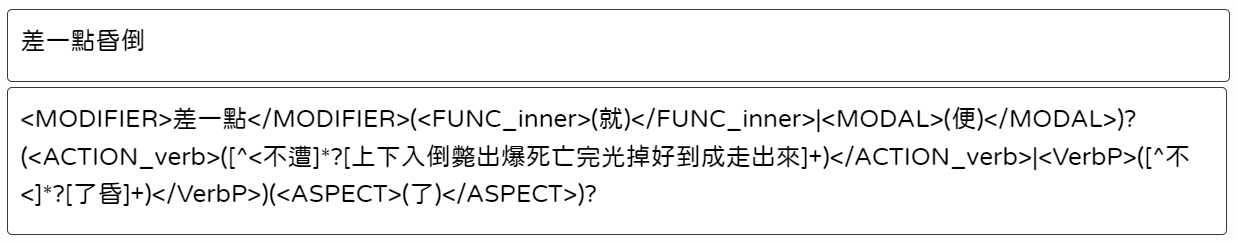
1. ) Loki的Regex輸出範例：



1. 以程式碼撰寫判斷依據和操作流程：

在 Loki Advanced 網頁工具上歸納並編輯正規表示式 (Regex) 以收斂句法結構態樣 (Syntactic Pattern) ，提高其壓縮比[[2]](#footnote-2)；此外，於正規表示式 (Regex) 中也可如例(19)所示，加入括號或以特殊語法撰寫也可以達到抽取特定詞彙參數或語法結構的目的。

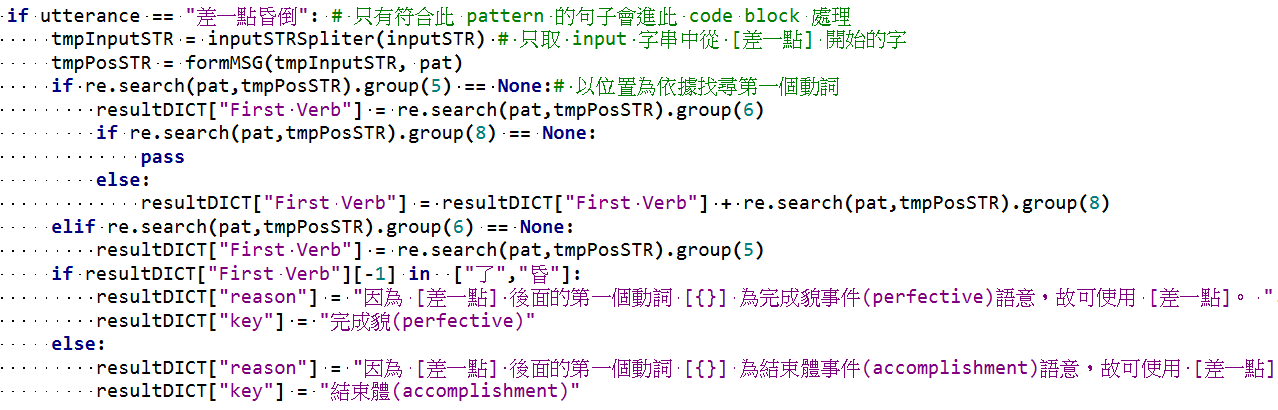
1. ) Regex的進一步編修：



根據對句法結構態樣 (Syntactic Pattern) 的觀察，「差一點」 後的動詞在構詞層面似乎帶有某種語意上的「錨點」。根據對語料的觀察在 Regex 中加入帶有「錨點」的動詞組詞尾，使其可抽取詞尾帶有「上」、「下」、「入」、「倒」、「斃」「出」、「爆」、「死」、「亡」、「完」、「光」、「掉」、「好」、「到」「成」、「走」、「出」、「來」的動詞及比對動詞前含有 「就」、「便」，動詞後含有動貌助詞「了」的句子，涵蓋性較編輯前有所提升。

接著將編輯過後的Loki 模型部屬完成後下載以 .py 檔案格式下載，Loki會自動完成比對句法結構態樣 (Syntactic Pattern) 的程式碼，使用者須在相關程式區塊中撰寫針對該句法結構態樣 (Syntactic Pattern) 的操作流程、判斷依據和解釋性文字，如例(20)所示。

1. ) XXX 輸入之語料符合「差一點昏倒」對應之Regex，即執行以下程式區塊。

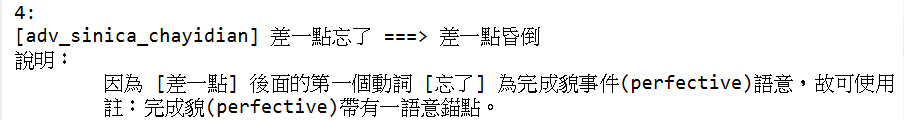


依據先前編輯的 Regex 位置抽取詞彙並加以組合得到「差一點」後的第一個動詞。經過觀察與歸納，依語意邏輯 (Semantic Logic) 設計歸納過後的判斷依據，抽取特定事件語意結構實現語意層面的分析，並將判斷依據加入程式區塊，使模型在運作過程中適時回傳判斷結果以利觀察修正。

1. 以模型大量批次處理語料：

將大量語料批次放進模型中自動驗證，觀察其結果並依結果提出假說和解釋，如例(21)。

1. ) 語感判斷程式化的範例：



在 Loki Advanced 中，來自中研院平衡語料庫之語料「差一點忘了」，句法結構態樣 (Syntactical Pattern) 與根據「差一點昏倒」設計之 Regex 相符，針對「差一點」後第一個動詞進行抽取判斷，得出其符合特定語意事件結構後給出解釋。根據上述處理程序，可根據大量語料驗證過後的結果，設計出針對近義詞分佈的語感判斷程式化系統。

## 研究規劃排程

|  |  |  |
| --- | --- | --- |
| 步驟 | 時間 | 執行任務 |
| 1 | 07.2023–08.2023 | 完成針對中研院平衡語料庫中，帶有「差一點」的語料處理觀察，句法結構態樣歸納及程式判斷撰寫。 |
| 2 | 08.2023–09.2023 | 完成針對中研院平衡語料庫中，帶有「差不多」的語料處理觀察，句法結構態樣歸納及程式判斷撰寫。 |
| 3 | 9.2023–10.2023 | 完成針對中研院平衡語料庫中，帶有「幾乎」的語料處理觀察，句法結構態樣歸納及程式判斷撰寫。 |
| 4 | 10.2023–11.2023 | 1. 排除錯誤及處理共現情形。 2. 彙整步驟1-3研究成果，撰寫摘要投稿會議發表。 |
| 5 | 11.2023–12.2023 | 1. 帶入大量語料驗證修正模型並驗證提出解釋。 2. 彙整步驟1-4研究成果，撰寫摘要投稿會議發表。 |
| 6 | 12.2023–02.2024 | 彙整步驟1-5研究成果，撰寫摘要投稿會議發表。 |

# 預期結果

透過觀察中研院平衡語料庫中含有「差一點」、「差不多」和「幾乎」的 2694 句語料，利用 Loki/ Articut 將語料轉換成句法結構態樣 (Syntactic Pattern) 進行分析，擬出針對其使用方式及特性，涵蓋性超越單一語料庫的解釋，並將過程以程式碼撰寫成可執行的程式模型，透過讓模型解析 Sketch Engine 中高達 30'000句的語料進行修正，提高解釋涵蓋性的同時留下具一致性及操作性的漢語近似副詞模型，此研究成果將提供未來漢語學習者自學的自動化應用工具，以解釋易產生混淆之漢語使用情形。

# 需要指導教授指導內容

1. 針對句法結構態樣 (Syntactic Pattern) 之分析
2. 針對語意邏輯 (Semantic Logic) 之分析歸納及解釋
3. 針對語意事件之分析及解釋
4. 針對語料共現情形的分析
5. 科學研究方法的落實及論文撰寫的細節
6. 學術倫理

# 參考文獻

Chuang Y.C. 2017. A comparison between “*chayidian*/*chadianr*/*chadian*” and “chabuduo” and their collocations. MA thesis. National Sun Yat-Sen University.

Jackendoff, Ray. 1974. *Introduction to the X-bar Convention*. Indiana University Linguistics Club.

Jackendoff, Ray. 1977. Constraints on Phrase Structure Rules, in P. W. Culicover, T. Wasow & A. Akmajian (eds.), *Formal Syntax*, Academic Press, New York, pp. 249–83.

Shen, J.-S. 1987. Chabuduo and chadianr. *Zhongguo* *Yuwen* [*Chinese* *Language*]: 442-456.

Shyu, S.-I. 2022. Scalar proximity and approximative adverbs in Mandarin Chinese. Talk given at 玉山語言學沙龍.

Wang, Wen-jet, Chen, Chia-jung, Lee, Chia-ming, Lai, Chien-yu, Lin, Hsin-hung. 2019. Articut: Chinese Word Segmentation and POS Tagging System [Computer program]. Version v.258, retrieved from [https://api.droidtown.co](https://api.droidtown.co/).

張家豪（2013）。論漢語中的終結點選擇。﹝碩士論文。國立臺灣師範大學﹞臺灣博碩士論文知識加值系統。 <https://hdl.handle.net/11296/qfg49a>

1. 中華民國發明證書號證書號 I750567。 [↑](#footnote-ref-1)
2. 以愈少的範例涵蓋愈多的測試句，壓縮比愈高，且模型效能愈好。 [↑](#footnote-ref-2)