## 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：

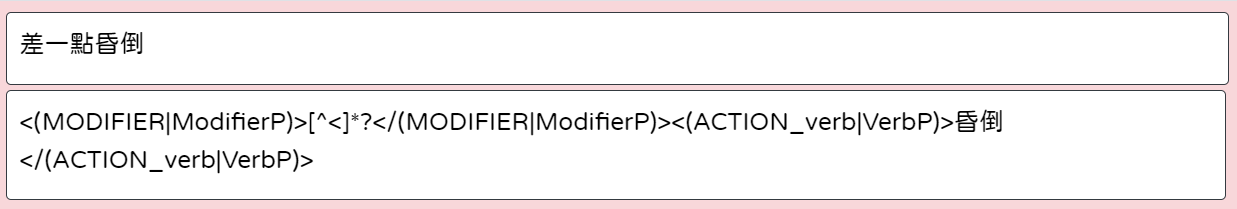
首先，使用Articut將語料進行中文斷詞 (CWS) 和詞性標記 (POS) 處理，而CWS和POS 賦予語料具一致性的操作符號，範例操作如(17)：

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)：

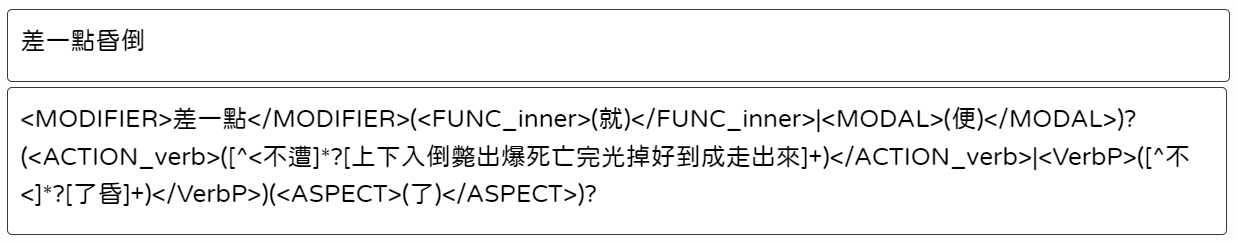
1. ) Loki的Regex輸出範例：



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[[1]](#footnote-1)；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。

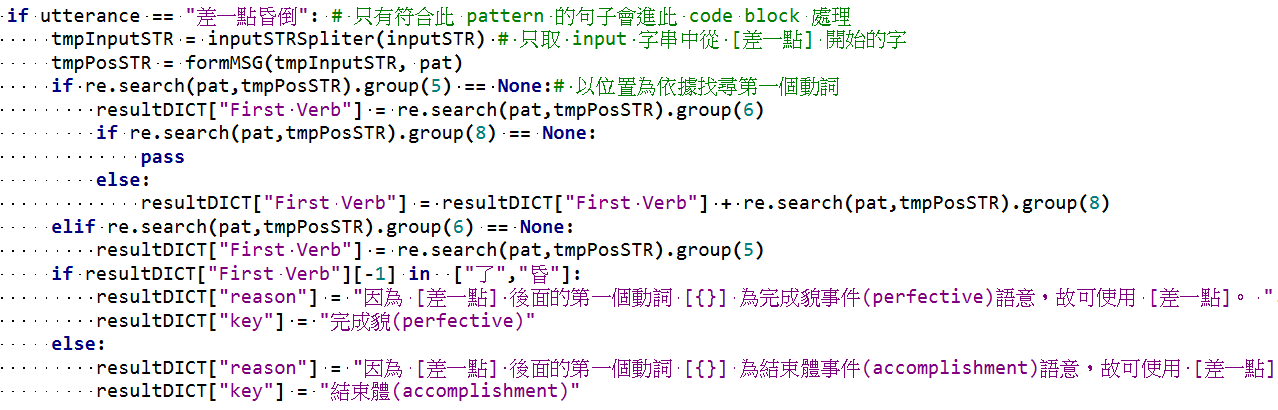
1. ) Regex的進一步編修：



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。

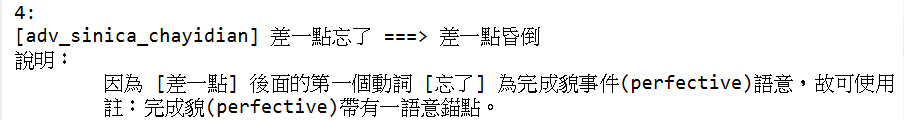


依據先前編輯的 Regex 位置抽取詞彙並加以組合得到「差一點」後的第一個動詞。經過觀察與歸納，依語意邏輯 (Semantic Logic) 設計歸納過後的判斷依據，抽取特定事件語意結構實現語意層面的分析，並將判斷依據加入程式區塊，使模型在運作過程中適時回傳判斷結果以利觀察修正。

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。

1. 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-1)