第二屆語料處理方法工作坊

語料處理方法入門

台大語言學研究所 謝舒凱

01

開幕

1. **Introduction**
2. Corpus Processing as Cooking Metaphor
3. Method/Methodology
4. Corpus and Social Practice

02

基本概念

| recapitualting the basic concepts

- 語料 (corpus data) 是語言研究的經驗基礎。
- 語料庫 (corpus; corpora) 本身不會直接提供 (認知) 語言學家要的東西。基本上，只提供字元字串的出現與否的訊息 (`information on the presence or absence of character strings`)。
  - 所謂字元字串可以包含各種語言單位：詞素、字詞、構式、標記等。
  - 與一般的資料庫的主要不同點：非結構性

基本素養

| some fundamentals

- 統計思維素養 (statistical thinking) 是必須的。
- 將知識與假說操作化 (`operationalization`) 與標記科學 (`annotation` science) 是必須的。
- 少動口，多動手。(hands-on linguistic project)

語料處理，處理什麼

| corpus data science workflow

- Collection
- Pre-processing: cleaning, tokenization, tagging (empowered by NLP)
- Index and query
- Exploratory data analysis (sta.significance testing and visualization)
- Computational Representation and (Machine Learning) Modeling
- Presentation (for scientific or commercial applications)

食說語料處理方法

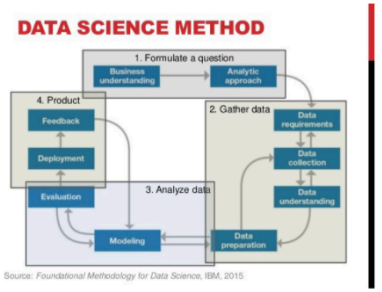
| corpus processing as cooking metaphor

Outline

1. Introduction
2. Corpus Processing as Cooking Metaphor
3. **Method/Methodology**
4. Corpus and Social Practice

資料(科學)處理方法

| data science method



語料處理方法

| corpus data processing methods

多了哪些？基本上可歸結於幾項關注訊息

- 頻率 **frequency**
- 散率 **dispersion**
- 共現的結構與呈現 `**coll((oc|ig)a|ostruc)tion**` and **concordance**
- 關鍵性 **keyness**

語言系統裡的交情問題

| **association**: from collocation to collostruction (Gries and Stefanowitsch, 2003;2005)

- 搭配現象 (collocation)
- 搭構 (`collostruction`) 是搭配 (`collocation`) 和構式 (`construction`) 的混搭新詞 (blending).
  - construction: form-meaning pairing that constitutes a basic unit in language.

共現的三種主要類型

- 簡單共現詞分析 (collexeme analysis)
- 顯著共現詞分析 (distinctive collexeme analysis)
- 共變共現詞分析：同一個構式中兩個位置的吸引程度。

語料處理方法

| corpus processing methods + Language Resources

- 在外部資源與 NLP 模型輔助下，尚可得到
  - 語言表徵 **representation**：以邏輯、網路、向量等方式表達。
  - 語言概念本體 **ontologies**

語料處理方法

| corpus processing methods and modelling

- 統計模型 statistical model：
  - 線性與非線性 linear and non-linear
- 語言模型 language model：
  - 統計機率與神經網路 statistical and neural

全域的語料視覺化

corpus data visualization [text visualization browser]

語料庫語言學的方法論

| methodological issues

從方法到方法論，是很少系統性觸及的一塊。舉例來說：

- 多大、多平衡、多具代表的語料，足以證成語言的專題分析？ What is the minimum corpus size for keyword and collocate testing? the bigger, the better?
- 語言資料的 取樣本質 不同，如何確保統計上的合理？
- 語言單位的 粒度 (granularity) 與穩定程度，如何反應在研究與應用上？

語料庫語言學的方法論

| exploratory data analysis

> *如果沒能清楚處理方法論的問題，我們就暫且把語料處理方法當成是【EDA】的一環(就好)，而不是用來驗證假說。*

語料標記、計算表徵與特徵學習

| corpus annotation, representation and feature learning

在語言科技的輔助下，還有幾個重要的主題需要了解

- 標記 (annotation)： 手作標記 (annotation) 與自動標記 (tagging)
- 特徵 (feture)：文本與語言特徵，機器學習中的特徵工程 (feature engineering)
- 表徵 (representation)：語言模型 (language model) 與嵌入 (embeddings)

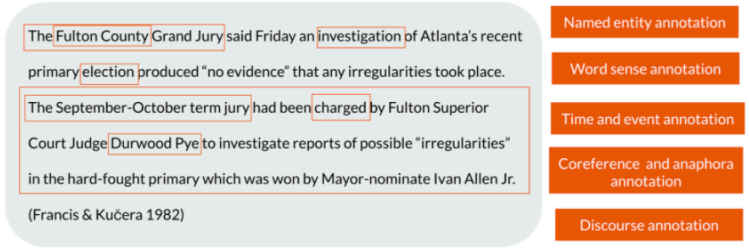
標記的層次

Levels of annotation

- **Linguistic levels** of corpus annotation:
  - Phonetic/Prosodic/POS/Syntactic/Semantic/Discoursal/Pragmatic/Stylistic
- **Paralinguistic levels** of corpus annotation:
  - Emotion/Affect/Personality
- **Conceptual levels**
  - ontological class

標記的粒度

(semantic) annotation granualrity



(photo courtesy: Andrea)

可以標的東西太多，依照任務來訂

More on Levels of Annotation

只要有(交際)功能，就有標記的空間。

- 類別 `class`、結構 `construction`、關係 `relation`、共指 `co-reference`、時間與事件 `temporal relations among events`、主題 `rhetorical topics` 等等。
- 語意語用、構詞句法、情緒評價、教學應用。

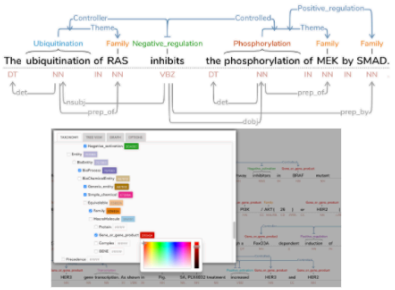
從單機走向服務

| corpus annotation framework

- GATE; ANNIS/Atomic: A web browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation.
- LOPE.anno; Tagtog; Label Studio
- RESTful Open Annotation

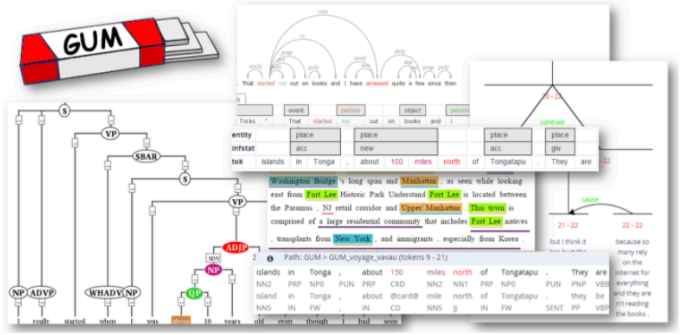
標記輔助工具日新月異

| text annotation graph



Multilayer annotation

個人是覺得有點走火入魔啦



標記的架構是人的工作

| Annotation scheme(s)

An annotation scheme should contain at least:

- a list of symbols used and an (operationalized) definition of their meaning.
- specification (explicit guidelines) of how decisions are made about how to attach information to linguistic units.

> *Clear enough to ensure a high degree of agreement if different annotators (also referred to as coders or raters) apply it to the same data.*

Linguistic Annotation Scheme

> *We can think of a linguistic annotation scheme as a comprehensive operational definition for a particular variable, with detailed instructions as to how the values of this variable should be assigned to linguistic data (in our case, corpus data, but annotation schemes are also needed to categorize experimentally elicited linguistic data).*

- 最理想是能夠相容 (compatible)、比較 (comparable)、交換 (exchangeable).

標記品質確保

| annotation quality

> *人會犯錯；更麻煩的是，如果沒有標準答案的時候怎麼辦*

- **Annotating** as `Interpretation`.
- Inconclusive and misleading results from linguistic analysis.

效度與信度

| validity vs. reliability

效度：指概念定義（Conceptual Definition）及操作化定義（Operational Definition）間是否契合

- The `validity` of annotation : whether the annotated categories are correct, but there is no "ground truth":
  - Linguistic categories are determined by human judgment
  - Consequence: we cannot measure correctness directly

效度與信度

| validity vs. reliability

信度：指可靠性或一致性。信度好的 scheme 在同樣或類似的條件下重複應用，可以得到一致或穩定的結果

- The `reliability` of annotation: whether human annotators consistently make same decisions → they have internalized the scheme.
  - Assumption: high reliability implies validity
  - How can reliability be determined?

各種標記工作的組合

- Each item is annotated by a single annotator, with random checks (≈ second annotation)
- Some of the items are annotated by two or more annotators
- Each item is annotated by two or more annotators - followed by reconciliation
- Each item is annotated by two or more annotators - followed by final decision by superannotator (expert)

In all cases, measure of reliability is to calculate the `coefficients of agreement`.

Coefficients of agreement

係數是什麼？

- The `Kappa coefficient` (κ, Cohen's kappa): a statistical measure of inter-rater reliability or agreement between two raters/annotators, ranging from 0 ('no agreement') to 1 ('complete agreement').

The Cohen's kappa

$$\kappa = \frac{P_o - Pe}{1 - Pe}$$

- P_o is the relative observed agreement between the raters (i.e. the percentage of cases where both raters have assigned the same category)
- P_e is the relative expected agreement (i.e. the percentage of cases where they should have agreed by chance).

指標解釋

| interpretation

κ	Level of agreement
0–0.20	None
0.21–0.39	Minimal
0.40–0.59	Weak
0.60–0.79	Moderate
0.80–0.90	Strong
> 0.90	Almost Perfect

情感表達挑戰了語言的結構觀

| affective texts/expression challenges

- Linguistic Units of affective expressions: the way we identify the expressive units of emotions will have influence on how researchers conceive their nature and their functioning.
- The formal treatment of language as prevalently assumed in linguistics, requires the sound and meaning-bearing linguistic units to be discretely distributed and governed by syntax. The analysis of emotional expressions under such framework will be restricted by the predefined grammatical boundary, like lexical or phrasal categories, etc.

以 ABSA (Aspect-based Sentiment Analysis) 為例

- `ABSA annotation`

Outline

1. Introduction
2. Corpus Processing as Cooking Metaphor
3. Method/Methodology
4. **Corpus and Social Practice**

語料庫與社會實踐

COVID19 corpus-based analytics

cf: Jupyter notebook

小結

hands-on corpus processing 的意義

- 語料處理是當今語言學專業的必修技能與知識，無需高估或貶抑，用心學了再說！(💪)
- 這個工作坊是上課同學邊學邊做的一個好榜樣（❤️）
- 如果時間不多，寧願多動手不動口 (😂)

願景

- 眾志成城：從自己的語料開始
- 加入 ptt 團隊 (`Public Taiwan Tensor`)： github and website
  - 台灣本土的語言語料庫建置、研究，與自然語言處理。
  - 永續精神與創意（原料與技術）。
  - 跨域跨界，結合文本、語音、手語、多模態。

讓我們都變成那個「沒有人」！

Reference

- Gries.S.(2020). *Ten lectures on corpus linguistics with R: Applications for usage-based and psycholinguistic research*. Brill.
- Stefanowitsch, A., & Gries, S.(2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2).
- Stefanowitsch, A., & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, 1(1).