吳小涵 / Hanna

- 台大計算語言學碩士
  - 阿諾標記有限公司
    - 艾斯移動

自然語言處理
Natural Language Processing

文本分類
Text Classifier

語言學/語意分析
Linguistic / Semantic Analysis

性別科技
Gender Tech.

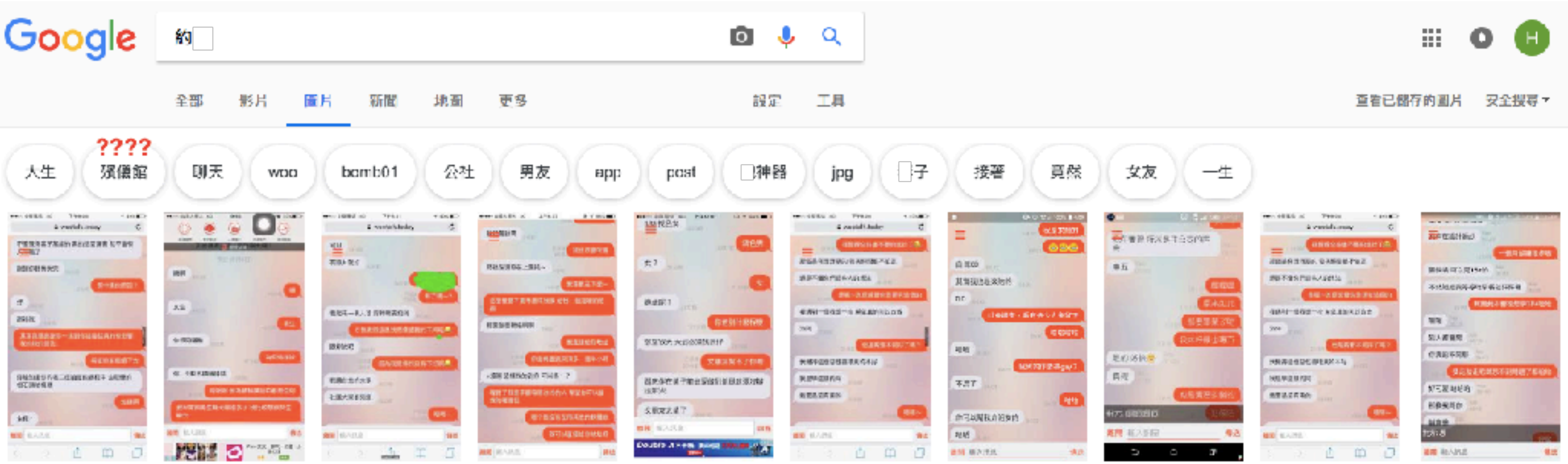
語言學/文本標記
Linguistic / Text Annotation

Projects

- 小姐聊色嗎

Data collection

- Google image
- 叔叔聊色嗎😏



Projects

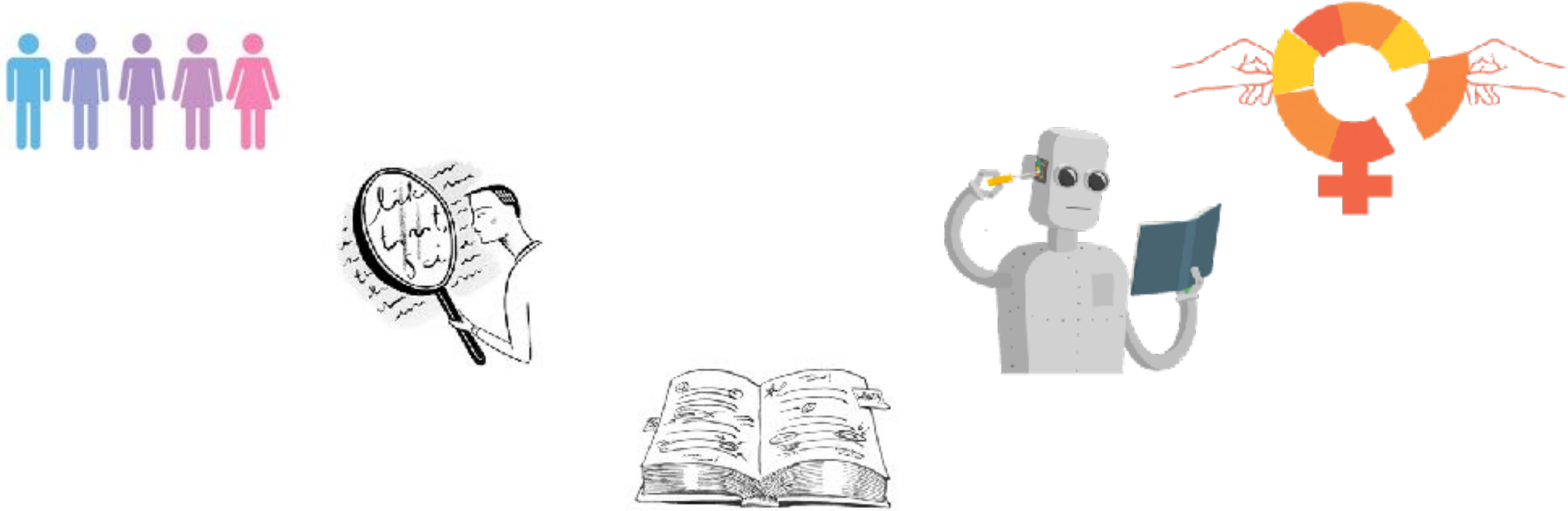
- 小姐聊色嗎

Data collection

- Google image
- 叔叔聊色嗎😏

Analysis

- 皮膚很好
- 皮膚很好餒~ 🏷
- 晚安
- 晚安安 🏷
- 晚亢



發現問題

尋找線索

線索標記

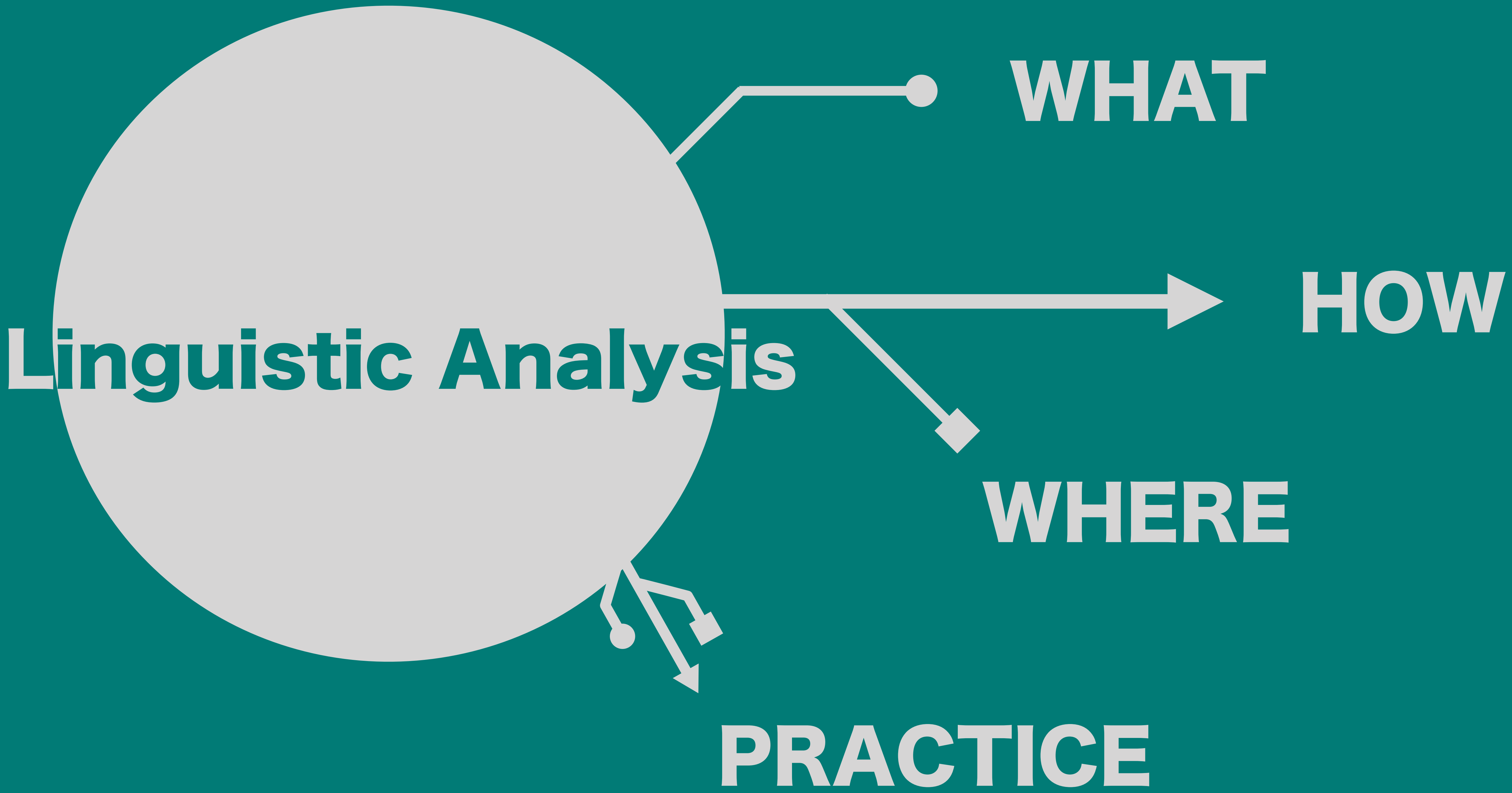
機器訓練

社會應用

語言學分析

語言學標記

收集資料



Linguistic Analysis

WHAT

HOW

WHERE

PRACTICE

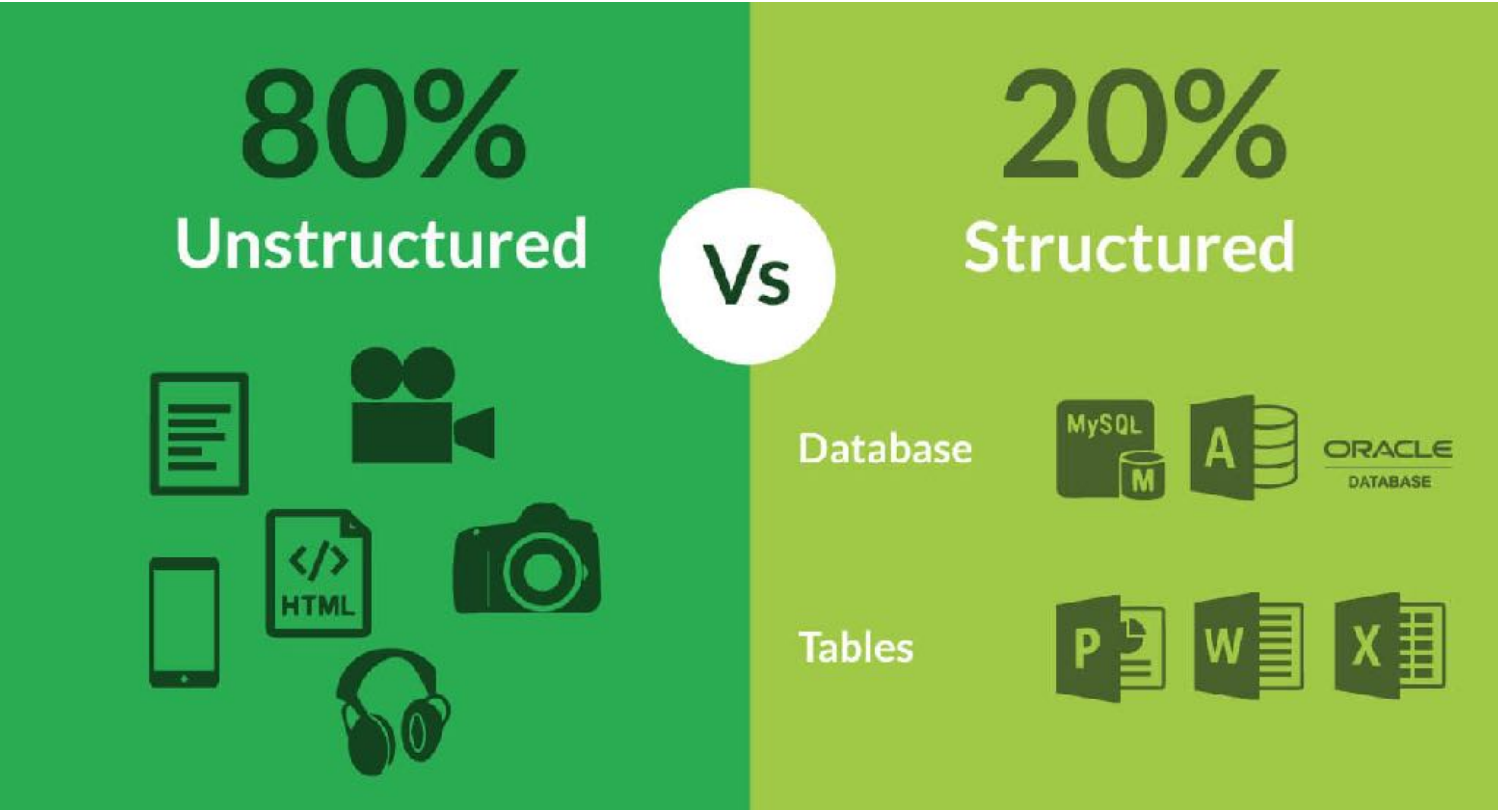
Linguistic Analysis

Huh ?

Social media

How to get insights from big data?

unstructured

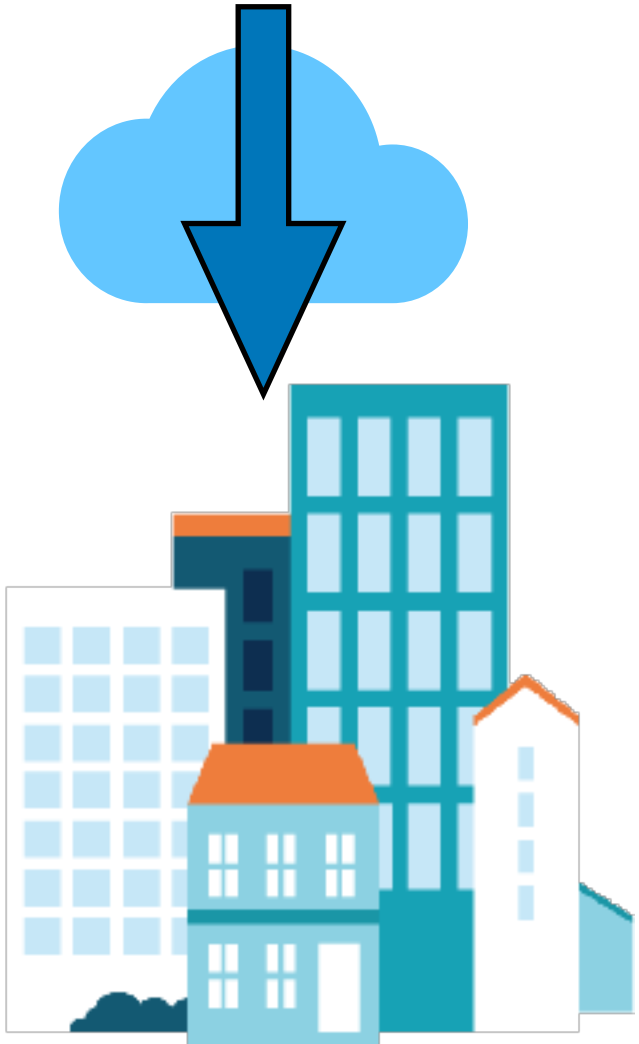


Adopted from: https://upxacademy.com/beginners-guide-to-big-data/

Social media

How to get insights from big data?

unstructured



Relies on text analysis solution that is accurate and detailed

Social media

twitter
You Tube
digg

How to get insights from big data?

unstructured

Umm… in..insight ?

💡 What the author intended to say

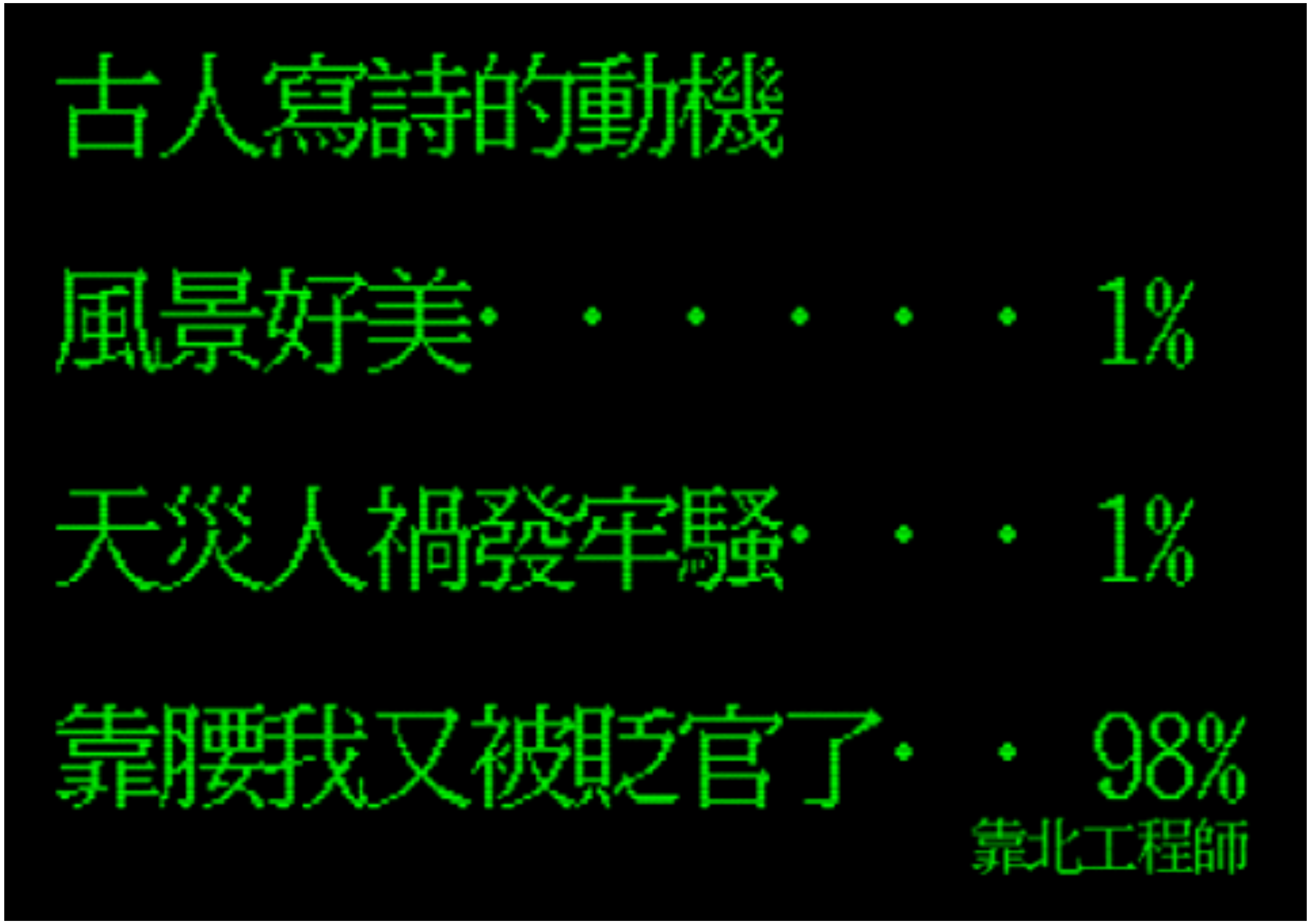
1. 以下兩個詞彙的詞性？

價格 (名詞)
包子 (名詞)

2. 以下兩句話作者想表達的心情?

價格 讚讚讚 (旅遊領域)
包子 讚讚讚 (科技領域)

To get the intended meaning of the given data



古人寫詩的動機

風景好美・・・・・・1%

天災人禍發牢騷・・・1%

靠腰我又被貶官了・・98%

靠北工程師

Linguistic Analysis

What to do?

NLP and Linguistic Analysis

- **Natural Language Processing (NLP)** is the ability for machine to understand and interpret human language the way it is written or spoken.

- The goal of NLP is to fill the gap how humans communicate and what the machine understands.

- From ***natural language*** to ***machine-readable language***

Before performing NLP, 3 levels of linguistic analysis should be considered:

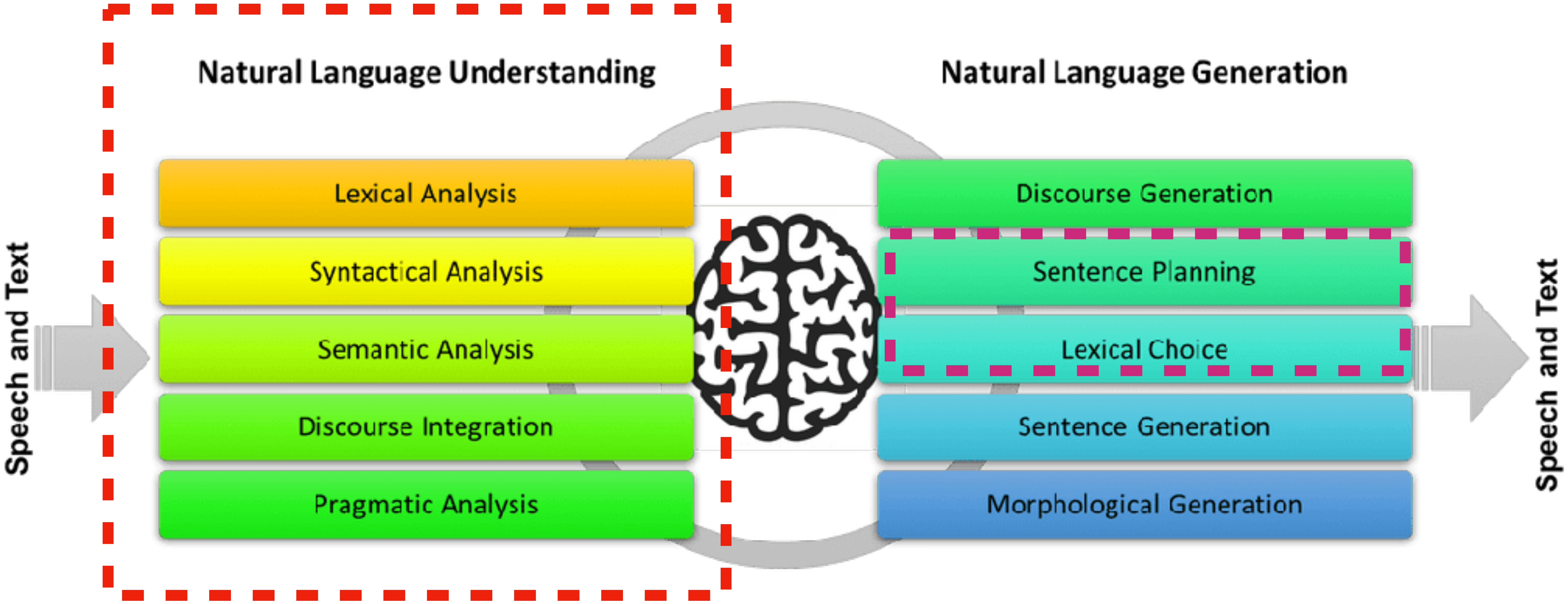
Syntax what part of given text is grammatically true

Semantics what is the meaning of given text

Pragmatics what is the purpose of given text

自然語言理解

自然語言生成



Speech and Text

Natural Language Understanding

Lexical Analysis

Syntactical Analysis

Semantic Analysis

Discourse Integration

Pragmatic Analysis

Natural Language Generation

Discourse Generation

Sentence Planning

Lexical Choice

Sentence Generation

Morphological Generation

Speech and Text

Natural Language Processing

Source:
FinTechXpert

To understand … what ?

- **Natural Language Understanding (NLU)** tries to understand the meaning of given text and tries to resolve the following ambiguities in natural language.

Lexical ambiguity	words have multiple meanings
Syntactic ambiguity	sentences have multiple parse trees
Semantic ambiguity	sentences have multiple meanings
Anaphoric ambiguity	word or phrase which is previously mentioned but has a different meaning

To understand … what ?

Lexical ambiguity

他真的很機車

Syntactic ambiguity

冬天：能穿多少穿多少； 夏天：能穿多少穿多少

||

Semantic ambiguity

女孩約的男孩遲到了有兩個原因: a. 睡過了. b. 睡過了

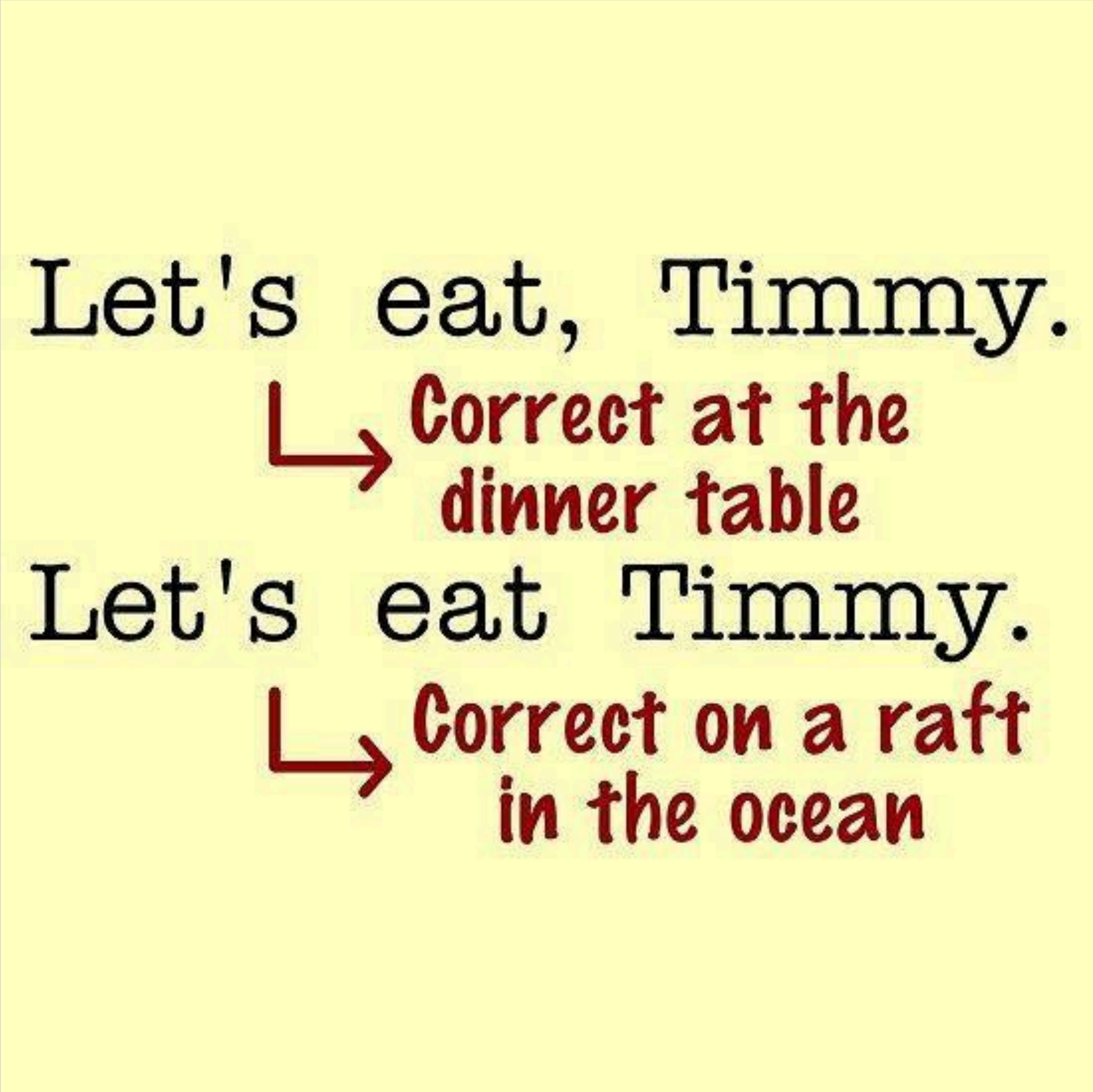
Anaphoric ambiguity

“Hanna invited Susan for a visit, and she gave her a good lunch."

(she = Hanna; her = Susan)

"Hanna invited Susan for a visit, but she told her she had to go to work"

(she = Susan; her = Hanna)



Let's eat, Timmy.

↳ Correct at the dinner table

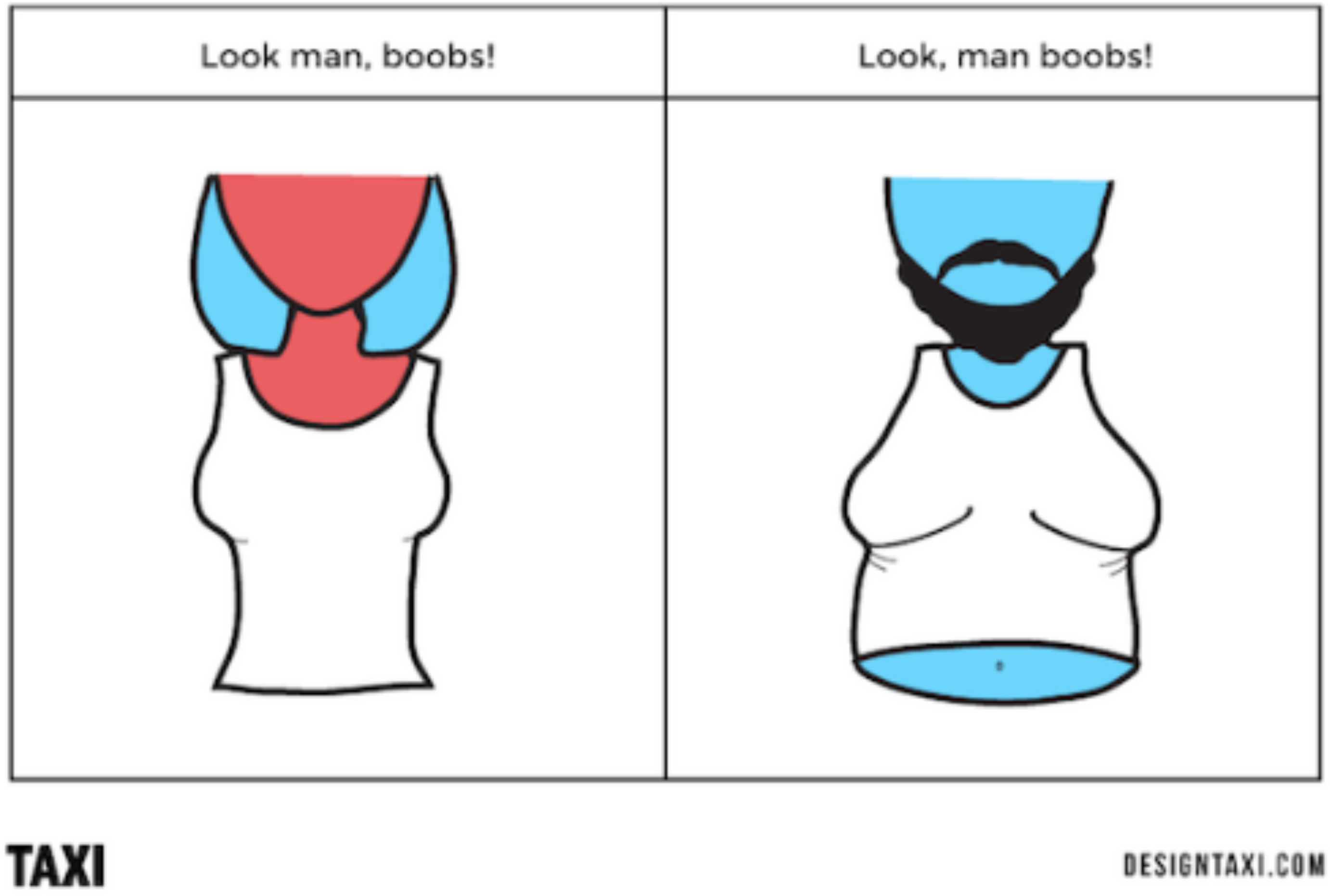
Let's eat Timmy.

↳ Correct on a raft in the ocean

Source: The language nerds.

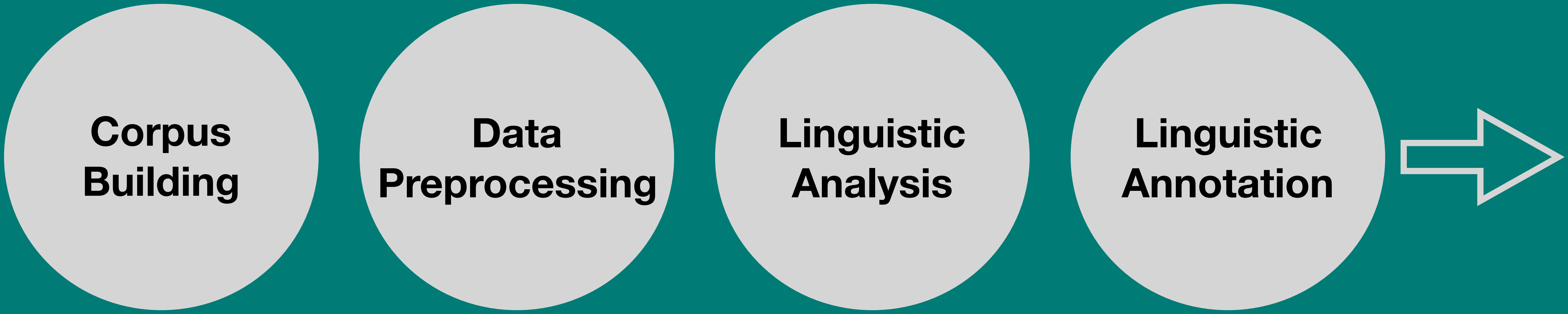
Google

the power of punctuation



Source: https://designtaxi.com/news/386071/Simple-Graphics-That-Show-The-Power-Of-Punctuation/

啊所以什麼時候才可以開始分析?



Corpus Building

Data Preprocessing

Linguistic Analysis

Linguistic Annotation

Web crawler

Data cleaning
Word segmentation

Extract the
INSIGHT
of collected data

Transfer language
to machine-readable
form

斷詞完就可以了啦!

Linguistic Analysis

How to do?

(好啦! 我們真的要來分析了!)

在語言的不同層次中，可以分析...?

Linguistic Level	What to analyze	Example
Lexical level	spelling error etymology	their → thier
Document level	topic distribution fake news	A報：腥羶色70%, 娛樂八卦20%, 政治10% O報：科普95%, 財經5%
Semantic level	semantic role	Mary sold the book to John (agent) (theme) (recipient)
Syntactic level	syntactic pattern	不就...而已嗎 (-) 連…也... (-)
Discourse level	turn-taking	overlapping/ interruption
Pragmatic level	speech act, intention	女友：「你不覺得她很漂亮嗎?」

.
.
.

Analysis also depends on your corpus

- Types of corpora

  Balanced, representative

  Monitor

  Parallel (translation)

  Comparable

  Diachronic

  Specialized

  Multi-media

  .
  .
  .

開始實作分析前，你需要準備的有：

- Regular Expression 的基本概念
- 見微知著的觀察力
- 思緒清晰的腦袋

先試試看下面的句子

內容農場怎麼用標題吸引點閱數

- 不分享是人嗎？醫生這番話，讓低頭族全都沉默了
- 2014最紅的10首歌！台灣人最愛的這首，竟是去年的
- 老婆把廚房整理成這樣，進門那一刻大家都震驚了！
- 超實用！大創39元「伸縮桿」竟有15種妙用！尤其第9項，解決所有女人的煩惱！趕快收藏起來！！

進階題來囉

透過短標題來分析文本情緒

- 考招研究 台大打臉清大：科學素養連小學生都不如 -
- 什麼時代了，都有支付寶了還吃狗肉，素質真是超高呢 -
- 昨晚又一個颱風形成 還好對台沒影響 +
- 下雨天逛outlet, 還好走道做得很好, 不會被雨淋到 +

除了那個「感覺」之外，為什麼你可以一讀就知道?

Linguistic Analysis

AND THEN…

WHERE ?

語言分析用來做什麼?

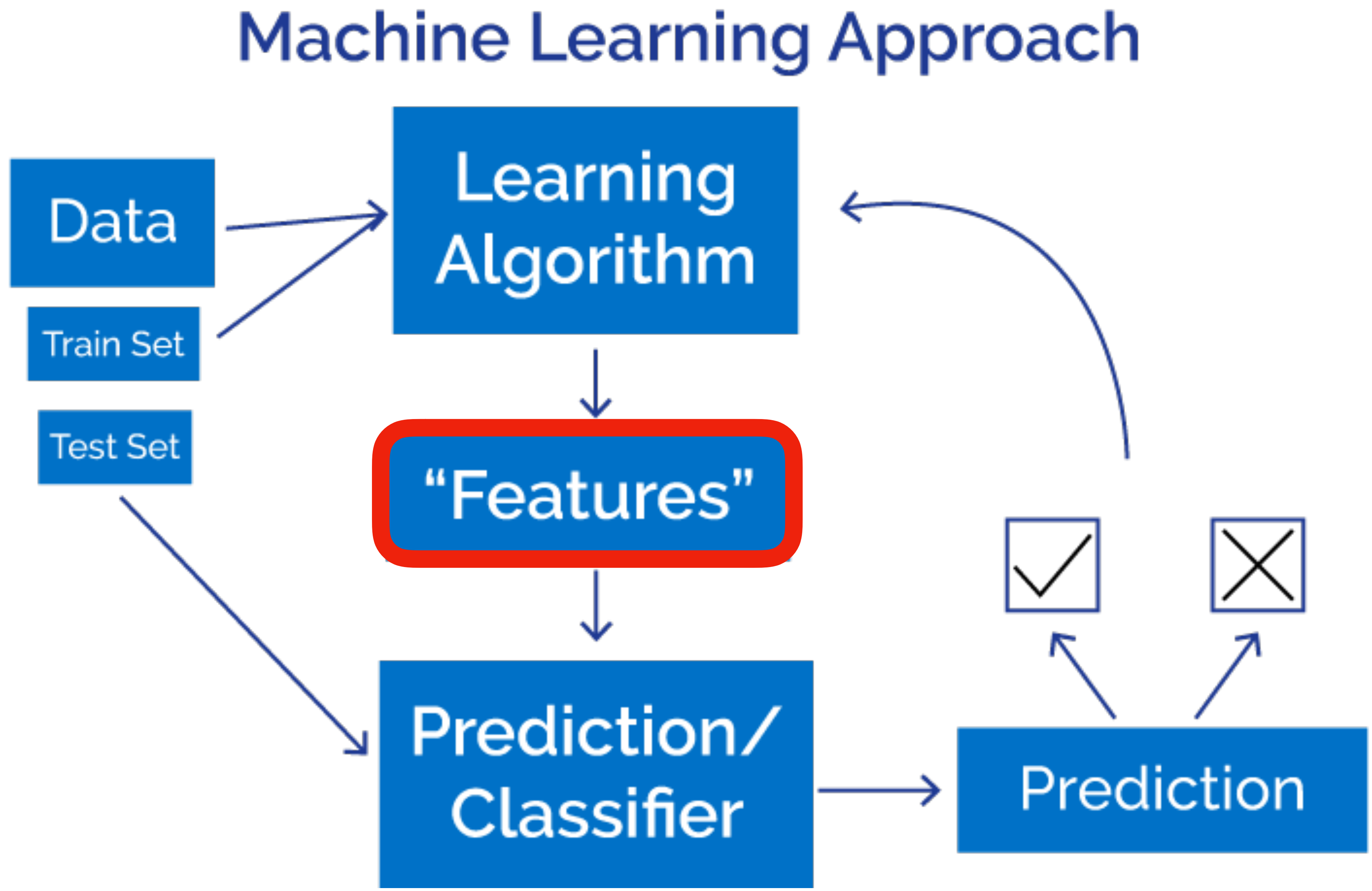
語言分析 + 網路文本 = 巨大的QA知識庫

舉例來說，常見的應用有：

1. Understand sentiment and emotion
2. Measure share of voice
3. Identify key topics, words, and phrases
4. Quantify purchase intent
5. Answer any question

https://www.crimsonhexagon.com/blog/how-text-analytics-works-for-social-media/

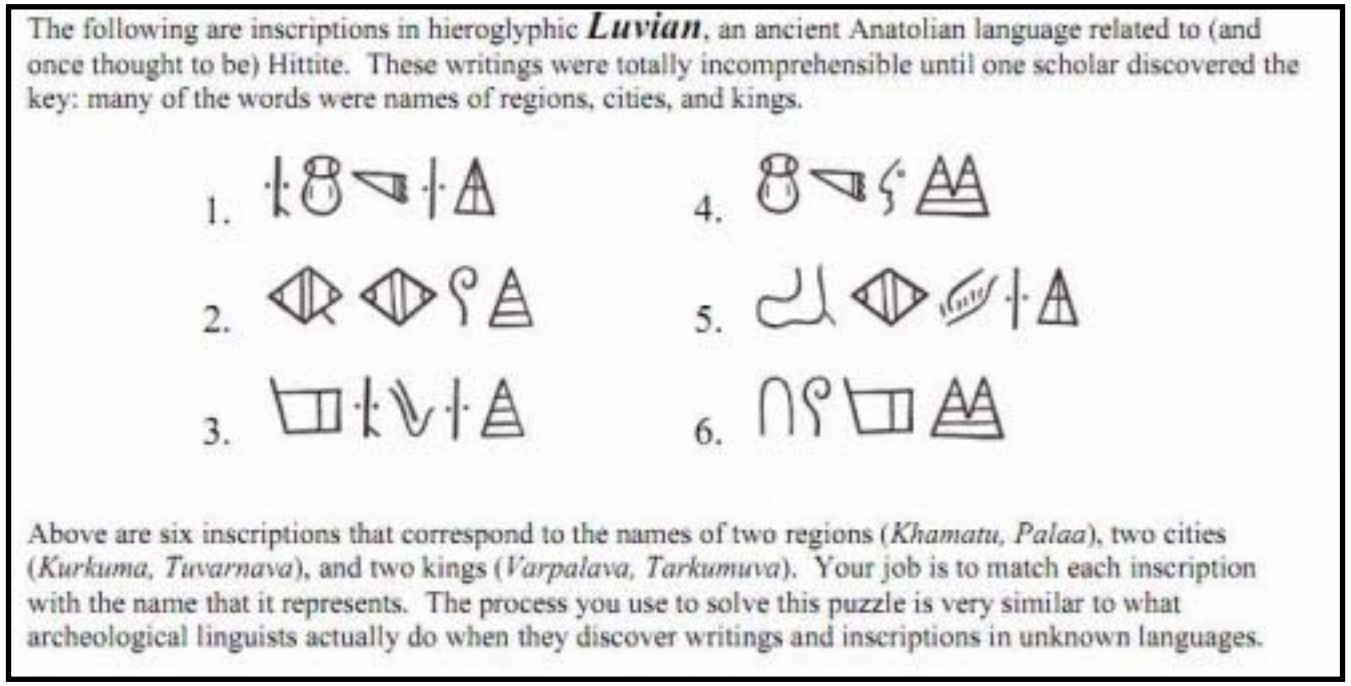
所以NLP的部份呢?



https://www.crimsonhexagon.com/blog/how-text-analytics-works-for-social-media/

回到語言分析本身的趣味性

分析不熟悉的「語言」



The following are inscriptions in hieroglyphic ***Luvian***, an ancient Anatolian language related to (and once thought to be) Hittite. These writings were totally incomprehensible until one scholar discovered the key: many of the words were names of regions, cities, and kings.

1.

2.

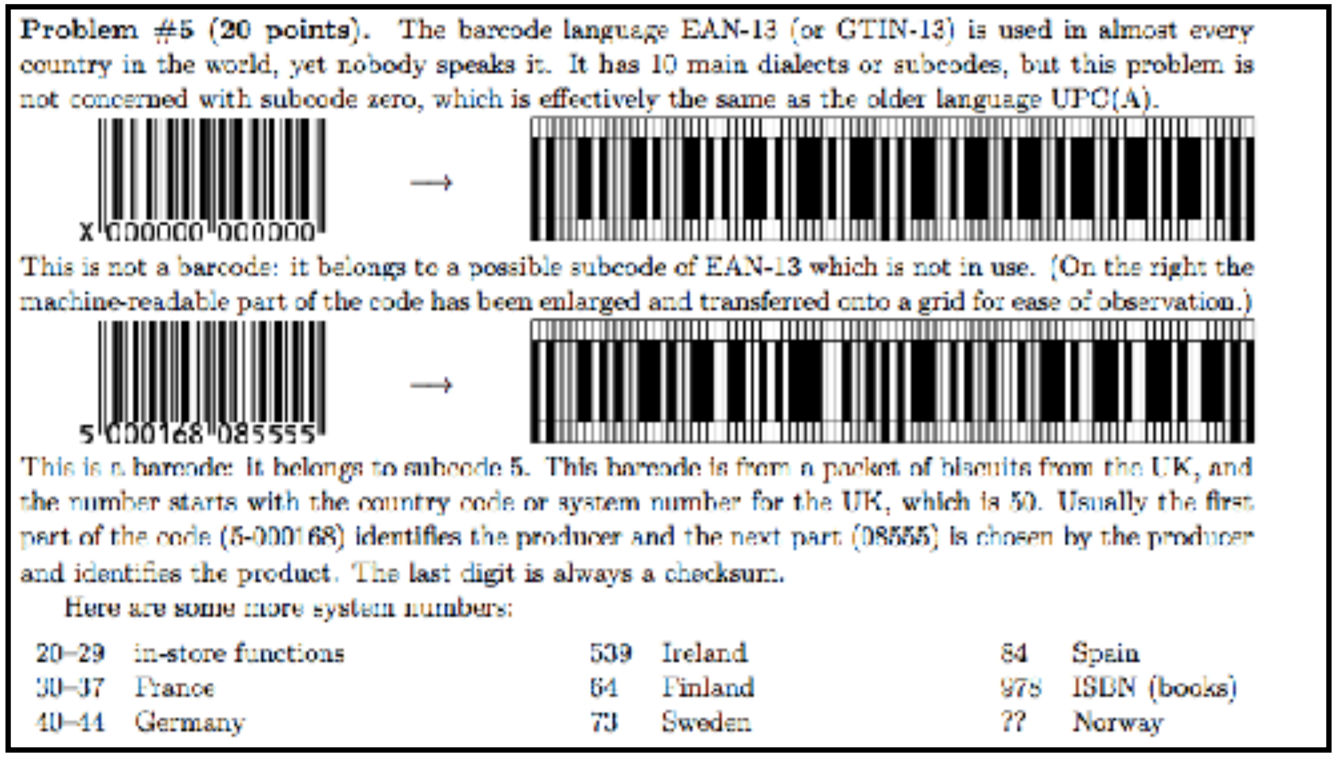
3.

4.

5.

6.

Above are six inscriptions that correspond to the names of two regions (*Khamatu, Palaa*), two cities (*Kurkuma, Tuvarnava*), and two kings (*Varpalava, Tarkumuva*). Your job is to match each inscription with the name that it represents. The process you use to solve this puzzle is very similar to what archeological linguists actually do when they discover writings and inscriptions in unknown languages.



Problem #5 (20 points). The barcode language EAN-13 (or GTIN-13) is used in almost every country in the world, yet nobody speaks it. It has 10 main dialects or subcodes, but this problem is not concerned with subcode zero, which is effectively the same as the older language UPC(A).

X 000000 000000 →

This is not a barcode: it belongs to a possible subcode of EAN-13 which is not in use. (On the right the machine-readable part of the code has been enlarged and transferred onto a grid for ease of observation.)

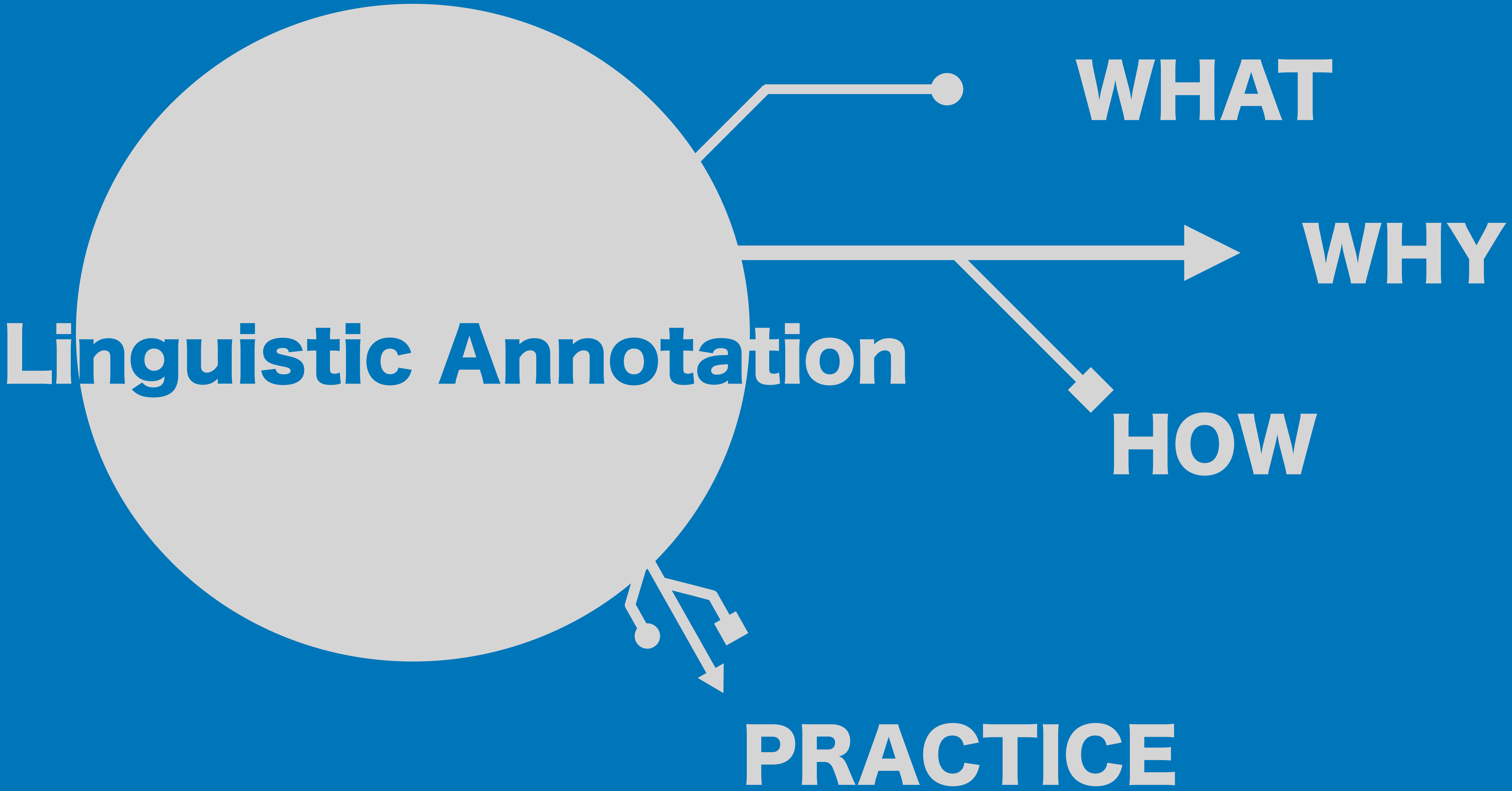
5 000168 085555 →

This is a barcode: it belongs to subcode 5. This barcode is from a packet of biscuits from the UK, and the number starts with the country code or system number for the UK, which is 50. Usually the first part of the code (5-000168) identifies the producer and the next part (08555) is chosen by the producer and identifies the product. The last digit is always a checksum.

Here are some more system numbers:

20–29	in-store functions	539	Ireland	84	Spain
30–37	France	64	Finland	978	ISBN (books)
40–44	Germany	73	Sweden	??	Norway

<ins>國際語言學奧林匹亞</ins>



Linguistic Annotation

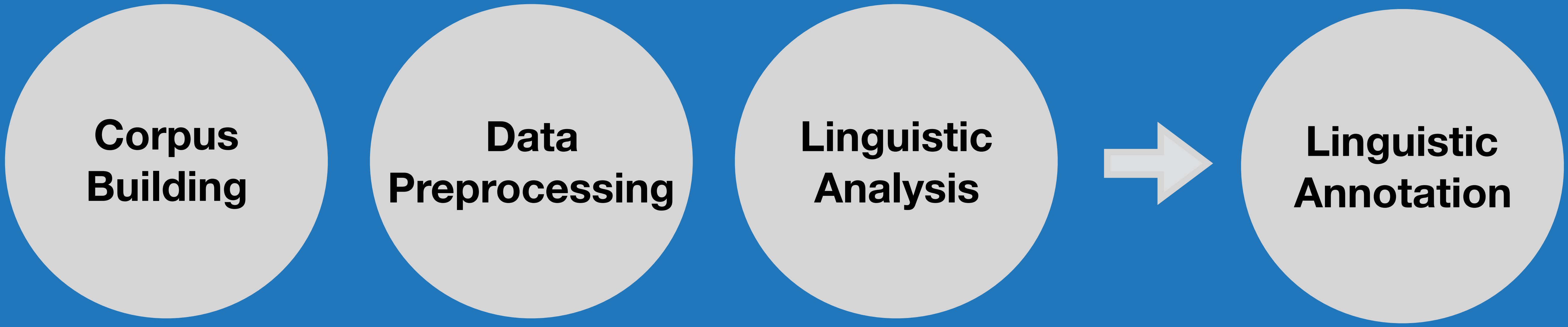
WHAT

WHY

HOW

PRACTICE

開始標記前...



Corpus Building

Data Preprocessing

Linguistic Analysis

Linguistic Annotation

網路爬蟲
自建語料庫

資料前處理
斷詞、斷句、除雜訊

擷取 INSIGHT

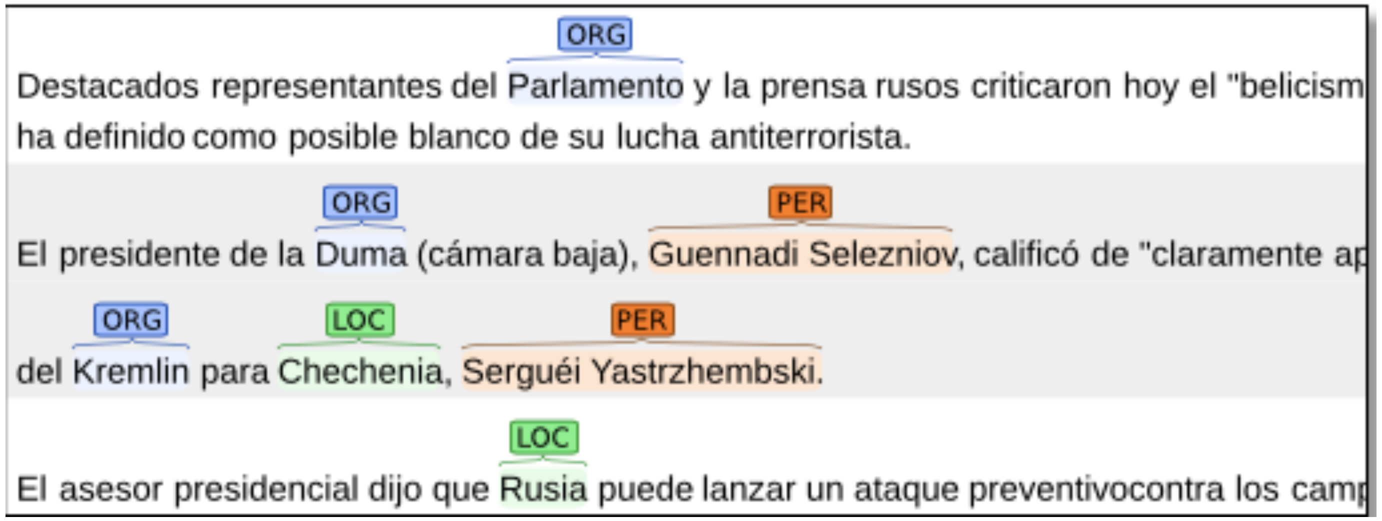
轉為
machine-readable form

什麼是語料庫標記？

- 將語言學資訊加註在語料庫中。
- *The practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data* (Leech 1997).

什麼是標記？

- **A note by way of explanation or comment added to a text or diagram.**



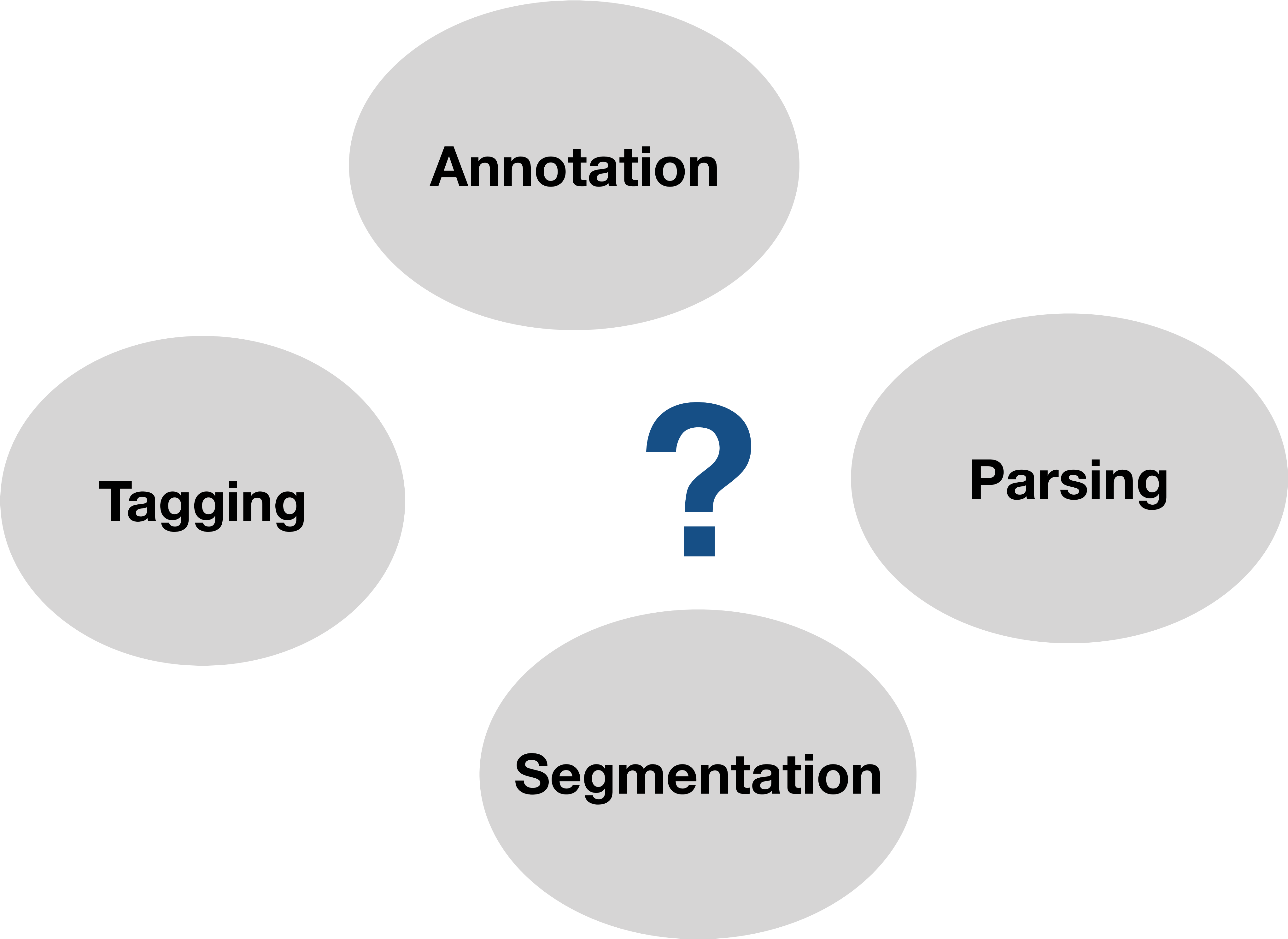
Destacados representantes del Parlamento y la prensa rusos criticaron hoy el "belicism
ha definido como posible blanco de su lucha antiterrorista.

El presidente de la Duma (cámara baja), Guennadi Selezniov, calificó de "claramente ap
del Kremlin para Chechenia, Serguéi Yastrzhembski.

El asesor presidencial dijo que Rusia puede lanzar un ataque preventivocontra los camp

http://brat.nlplab.org/examples.html

什麼是標記？



Annotation

Tagging

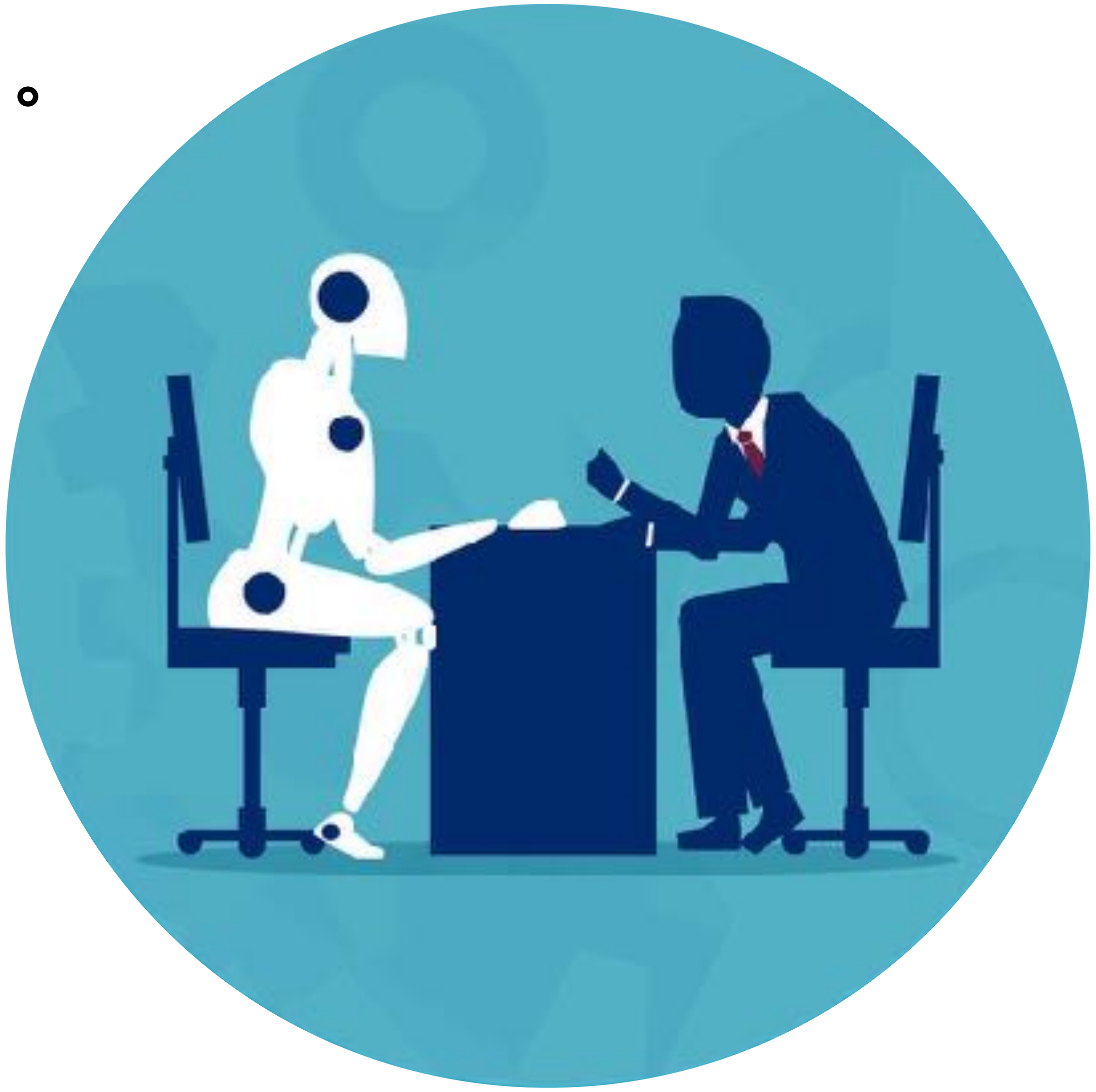
?

Parsing

Segmentation

為什麼需要標記？

- 將人類語言及知識轉成機器可讀取的語言資訊。
- 給予語料庫不同的資訊價值。



標記的層次

- 發展完善，高度自動化的層次

詞性標注 Part of Speech (PoS) tagging

句法剖析 Syntactic Parsing

斷句/斷詞 Sentence/ word Segmentation

標記的層次

- 需要更多人工語言分析介入的層次

詞彙標記 Lexical Annotation

語意標記 Semantic Annotation

語用標記 Pragmatic Annotation

言談標記 Discourse Annotation

標記準則

- 標記資訊可與文本資料切分，及可回復至未標記前的文本樣貌。
- 明確的標記過程及標記者。
- 標記資訊結果不是”真理”，而是能善用的資源。
- 標記架構應以現有理論為基礎。
- 沒有任何標記架構能一步到位，多方面考量後訂定。

標記架構

- 要標什麼
- 要標多少
- 怎麼標
- 用什麼標
- 標哪裡
- 標記規則訂定

語料庫標記層次

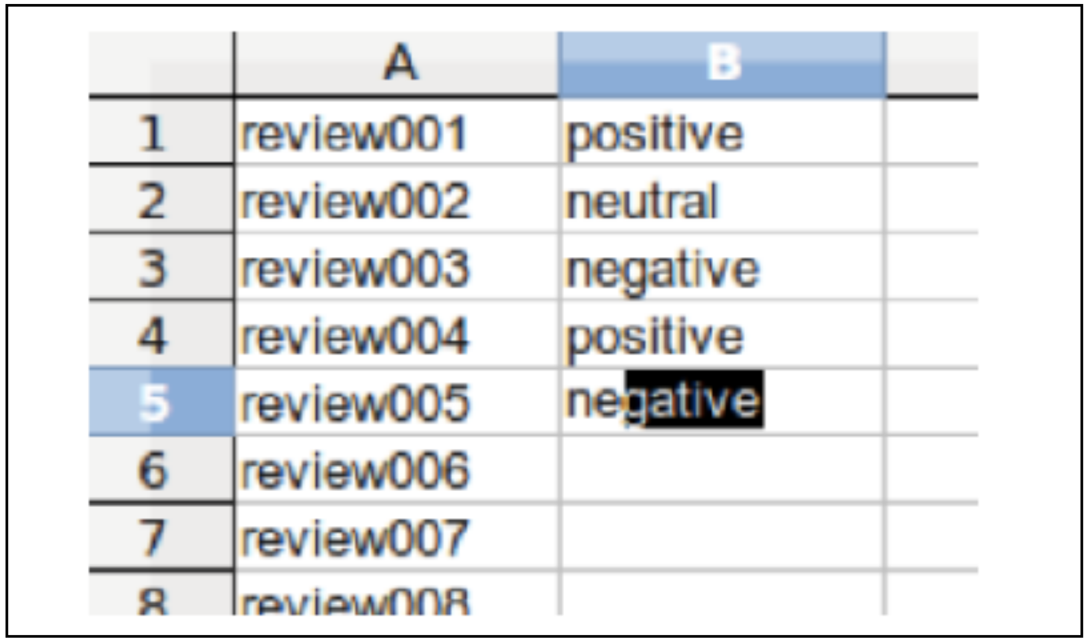
- **Metadata (文檔號碼、作者資訊、出版日期等等)**
- **Textual markup (XML 格式)**
- **Linguistic annotation (語料標記內容)**

語言學標記

- 詞義 word sense
- 句法歧異 syntactic ambiguity
- 情緒 Sentiment
- 語意 Semantic
  - 諷刺性
  - 評價語言
  - 立場
- 語用 pragmatic

標記格式

手工型



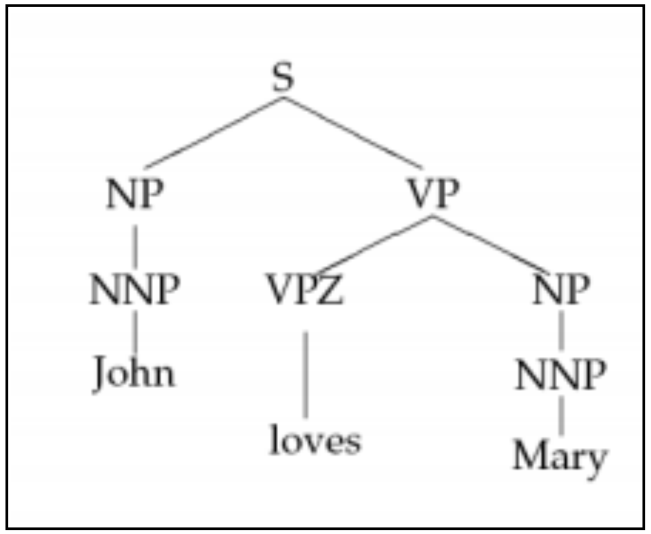
	A	B
1	review001	positive
2	review002	neutral
3	review003	negative
4	review004	positive
5	review005	negative
6	review006	
7	review007	
8	review008	

傳統型

John loves Mary.

⇓

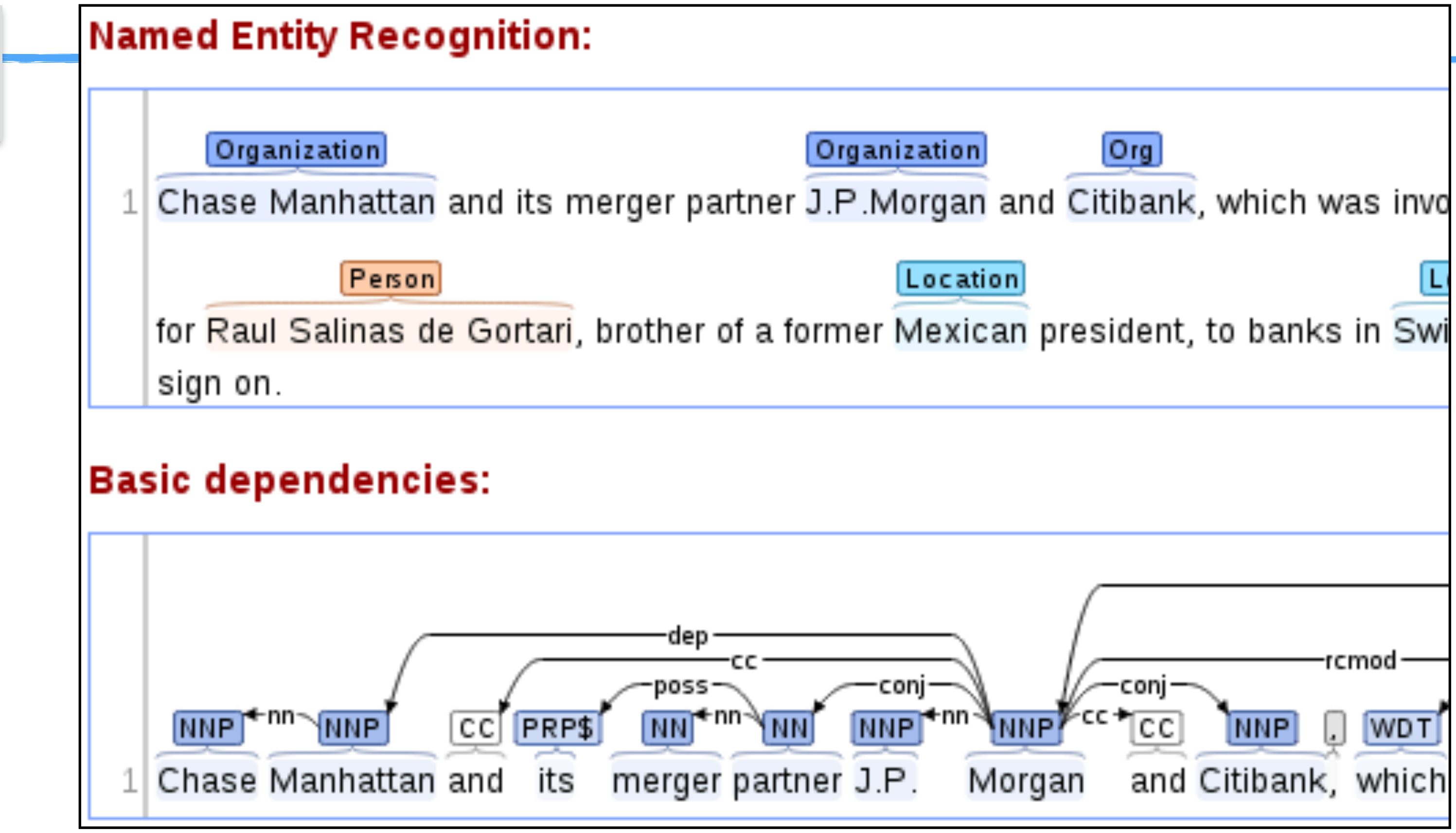
```
(S (NP (NNP John))
   (VP (VPZ loves)
        (NP (NNP Mary)))
   (. . ))
```



Source: Natural Language Annotation for Machine Learning: A guide to corpus-building for applications
O'Reilly Media, Inc.
Source: http://brat.nlplab.org/examples.html

標記格式

華麗型



Named Entity Recognition:

1 Chase Manhattan and its merger partner J.P.Morgan and Citibank, which was invo

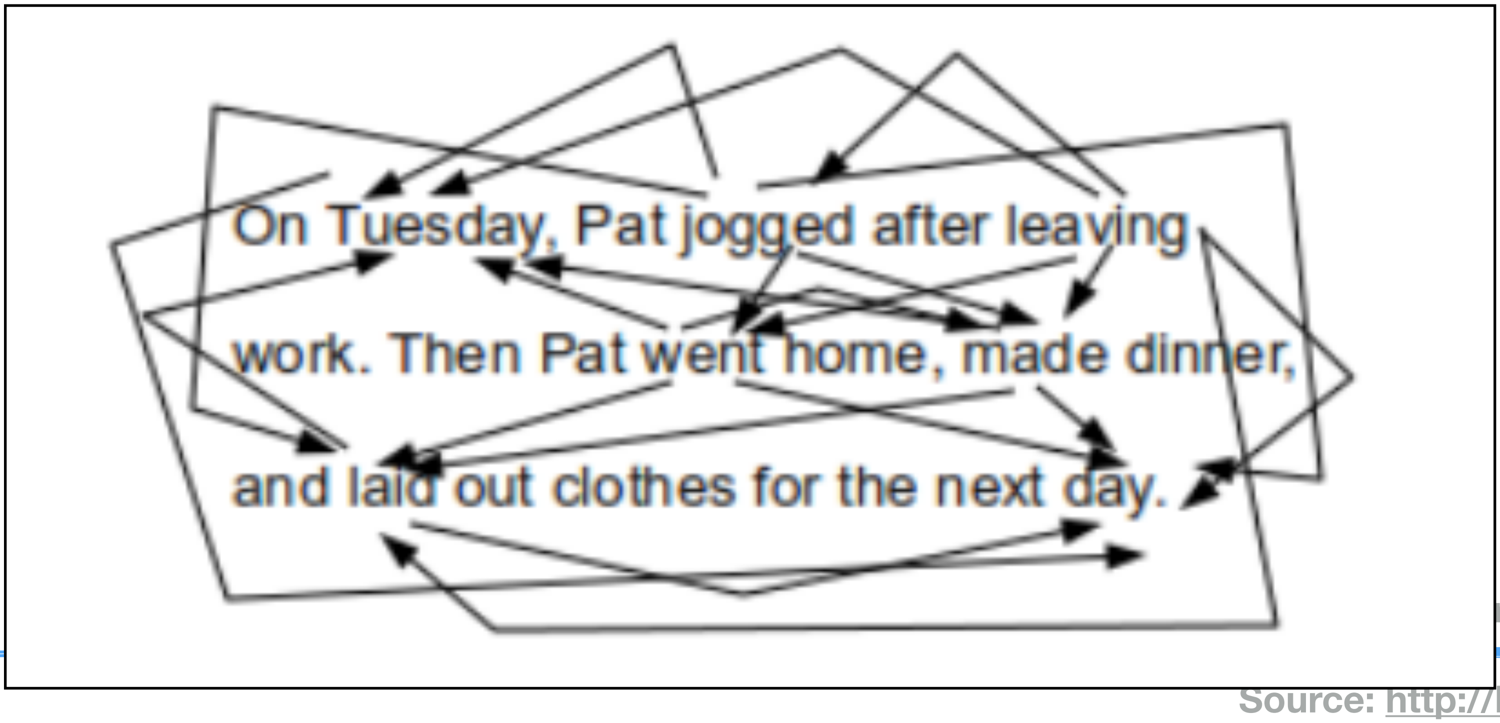
for Raul Salinas de Gortari, brother of a former Mexican president, to banks in Swi

sign on.

Basic dependencies:

1 Chase Manhattan and its merger partner J.P. Morgan and Citibank, which

瘋狂型



On Tuesday, Pat jogged after leaving work. Then Pat went home, made dinner, and laid out clothes for the next day.

rpus-building for applications
O'Reilly Media, Inc.
Source: http://brat.nlplab.org/examples.html

符碼意義

- **決定標記的符碼，保持一致性。**

情緒強度	scale 1-5	0,1,2,3,4
情緒強度	極度	+, -
情緒呈現	有無	YES,NO 0,1
	單位劃分	<情緒詞>高興</情緒詞>

語料庫標記難處

- 歧異的處理
- 遇到未知詞或新詞
- 語言使用變遷的影響

標記工具資源

TagAnt: POS Tagger

USAS Web Tagger (POS and XML)

UCREL Chinese Semantic Tagger

中研院平衡語料庫(詞類)

- 中研院版本的tagset

http://lopen.linguistics.ntu.edu.tw/lope.anno/

http://corpus-tools.org/annis/

自然語言處理(NLP)的標記資源

NLTK

Python 3.5

```
# Using nltk for English text annotation
>>>from nltk import pos_tag, word_tokenize
>>>pos_tag(word_tokenize(‘This is an English sentence.’))
```

```
# Using nltk for Chinese text annotation
>>>from nltk.corpus import sinica_treebank
>>>sinica_treebank.words()[10:15]
>>>sinica_treebank.tagged_words()[10:15]
>>>sinica_treebank.sents()[13]
>>>sinica_treebank.parsed_sents()[13].draw()
```

分析標記一次來

語言學分析 & 標記實作

- 資料準備
- 語言學分析可能的角度
- 標記架構、準則、一致性確立
- 進行標記
- Lopotator 標記工具平台

Linguistic Annotation

語料庫標記實作

PRACTICE

Prerequisites

- regular expression 基本概念
- 觀察力
- 思緒清晰
- 維持標記準則

Sentiment Annotation

Download your data here: https://reurl.cc/nnrZd

資料夾內的檔案：

readme.txt
positive_words.txt
negative_words.txt
new_sample.txt
sentiment_annotation.ipynb
sentiment_annotation.py

Sentiment Annotation

在開始之前

```
# Install packages that are needed for our annotation practice
pip install csv
pip install pandas
pip install re
```