2019 0416 第九週
蒐集詞彙語意研究資料
——網路爬蟲

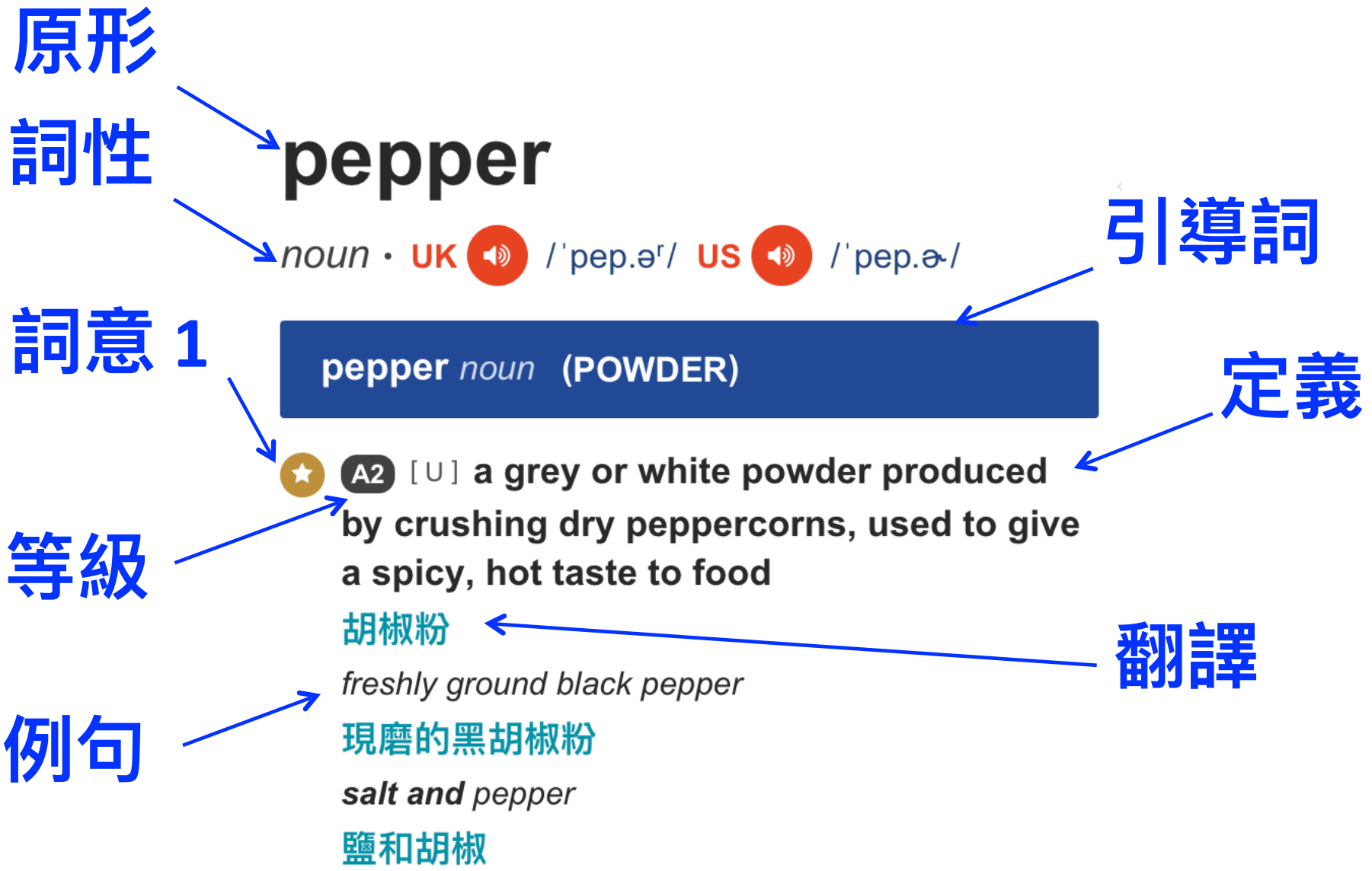
Pepper 有幾種意思?

salt and pepper
roast red pepper in the oven

會先想到：查劍橋英中辭典

https://dictionary.cambridge.org/us/dictionary/english-chinese-traditional/pepper

詞, 原形, 詞意, 定義


原形
詞性
pepper
noun · UK /ˈpep.əʳ/ US /ˈpep.ɚ/
引導詞
詞意 1
pepper noun (POWDER)
定義
A2 [U] a grey or white powder produced by crushing dry peppercorns, used to give a spicy, hot taste to food
等級
胡椒粉
翻譯
freshly ground black pepper
例句
現磨的黑胡椒粉
salt and pepper
鹽和胡椒

詞, 原形, 詞意, 定義

詞意 2

pepper *noun* **(VEGETABLE)**

B1 ALSO sweet pepper, US ALSO bell pepper [C] **a vegetable that is usually green, red, or yellow, has a rounded shape, and is hollow with seeds in the middle**

定義

甜椒，燈籠椒

翻譯

a red/green pepper

紅 / 青椒

等級

Peppers are usually cooked with other vegetables or eaten raw in salads.

甜椒通常與其他蔬菜一同烹調，或放在沙拉裡生吃。

例句

***Red** peppers are ideal for roasting in the oven.*

紅椒很適合用烤箱烘烤來吃。

HTML 基本概念 https://en.wikipedia.org/wiki/HTML

- Texts + nested elements
  - start tag, end tag
- Start tag may contain
  - attributes
    - value (one value (sometimes more) per attribute

```
<tag attribute1="value1" attribute2="value2">
<tag attribute1="value1 value2">
```

    - Use `class` attribute to classify similar elements (semantics)
- Example

```
<p>Paragraph 1</p> <p>Paragraph 2</p>

<a href="https://www.wikipedia.org/">A link to Wikipedia!</a>

<input type="text" /> <!-- This is for text input -->
<input type="file" /> <!-- This is for uploading files -->
<input type="checkbox" /> <!-- This is for checkboxes -->
```

相對網頁的格式 (1)

headword
pepper

pepper

noun · UK /ˈpep.əʳ/ US /ˈpep.ɚ/

headword + pos + guideword
pepper noun
(POWDER)

pepper *noun* **(POWDER)**

epp-xref: A2
gram: [U]

A2 [U]

def + trans
<b class="def">a grey or white powder produced by crushing dry peppercorns, used to give a spicy, hot taste to food.胡椒粉

A2 [U] **a grey or white powder produced by crushing dry peppercorns, used to give a spicy, hot taste to food**

胡椒粉

eg and trans
#1: freshly ground black pepper
現磨的黑胡椒粉

#2: salt and pepper
鹽和胡椒

freshly ground black pepper

現磨的黑胡椒粉

***salt and** pepper*

鹽和胡椒

相對網頁的格式 (2)

headword + pos + guideword

pepper noun (VEGETABLE)

epp-xref B1

ALSO sweet pepper, US ALSO bell pepper

usage+v also sweet pepper, US also bell pepper

gram [C] [C]

vegetable that is usually green, red, or yellow, has a rounded shape, and is hollow with seeds in the middle

甜椒，燈籠椒

def + trans

<b class="def"> a vegetable that is usually green, red, or yellow, has a rounded shape, and is hollow with seeds in the middle 甜椒，燈籠椒

a red/green pepper

紅 / 青椒

Peppers are usually cooked with other vegetables or eaten raw in salads.

甜椒通常與其他蔬菜一同烹調，或放在沙拉裡生吃。

***Red** peppers are ideal for roasting in the oven.*

紅椒很適合用烤箱烘烤來吃。

eg + trans

#1: a red/green pepper
紅 / 青椒

#2: Peppers are usually cooked with other vegetables or eaten raw in salads. 甜椒通常與其他蔬菜一同烹調，或放在沙拉裡生吃。

#3: Red peppers are ideal for roasting in the oven.
紅椒很適合用烤箱烘烤來吃。

相關模組

- 下載網頁模組
  - import urllib
  - page = urllib.request.urlopen(<url>)
- 分析網頁模組
  - from bs4 import BeautifulSoup
  - soup = BeautifulSoup(page_source, ‘html.parser’)
  - for div in soup.find_all([<tag>, …])
  - check div[<attribute>]

用下載資料研究詞彙語意解岐

- 利用詞典英文、中文定義詞、引導分類詞，做語意解岐
  - 輸入：句子 (問題詞＋文脈詞） presidential chopper
  - 輸出：詞典的語意條目 AIRCRAFT; TOOL; MOTORCYCLE
  - **chopper**

noun · UK /ˈtʃɒp.əʳ/ US /ˈtʃɑː.pəʳ/

chopper *noun* (AIRCRAFT)

[C] INFORMAL FOR **helicopter**
直升機（helicopter的非正式說法）

chopper *noun* (TOOL)

[C] **a small axe held in one hand**
小斧頭

chopper *noun* (MOTORCYCLE)

a type of motorcycle with tall handlebars (= parts that you hold) and a long front end
（高車把、長前叉的）美式機車

a type of bicycle that is designed to look similar to a chopper motorcycle
設計得看上去像是美式機車的自行車

研究議題 (1)

- 利用詞典英文、中文定義詞、引導分類詞，做語意解岐
  - 輸入：presidential chopper
  - 輸出：AIRCRAFT; TOOL; MOTORCYCLE
  - 建立關係： <文脈詞, 語意定義詞+引導分類詞>
    - (presidential, #1) —> (presidential, AIRCRAFT)
      (presidential, #2) —> (presidential, TOOL)
      (presidential, #3) —> (presidential, MOTORCYCLE)
  - 研究方法
    - Bayesian models (supervised, unsupervised)
    - EM models
    - Graph models

研究議題 (2)

- 自動調整語意條目的引導詞分類
  - 問題：引導詞分類不一致 —> 解答：自動化的多重分類
    - pepper#1 = POWDER —> #spice, #food
      salt#1 = FOOD —> #spice, #food
      saffron#1 = SPICE —> #spice, #food
    - chopper#1 = AIRCRAFT —> #aircraft
  - 問題：非歧義詞缺乏引導詞分類 --> 解答：自動分類
    - sugar#1 = ? —> #spice, #food
    - airplane#1 = ? —> #aircraft
      helicopter#1 = ? —> #aircraft
      aircraft#1 = ? —> #aircraft
    -

參考文獻

- Bayesian models (unsupervised): Yarowsky (1992)
  - Word-sense disambiguation using statistical models of Roget's categories trained on large corpora
- EM-like models: Yarowsky (1995)
  - Unsupervised word sense disambiguation rivaling supervised methods
- Graph models: Agirre et al. (2014)
  - Random walks for knowledge-based word sense disambiguation
- Topic models: Chaplot and Salakhutdinov (2018)
  - Knowledge-Based Word Sense Disambiguation Using Topic Models
- Neural models: Yuan et al. (2016)
  - Semi-supervised word sense disambiguation with neural models

研究議題 (cont.)

- 利用 word embeddings