語詞

Gú-sû

Word

1. 詞的定義 Sû ê tēng-gī

語意的最小單位

Gú-ì siōng sòe ê tan-ūi

分詞規範是計算機處理的根據

Hun-sû kui-hōan sī kè-sǹg-ki chhú-lí ê kun-kù

2. 詞型 / 詞次 Sû-hêng / sû-chhù

詞型 **sû-hêng (word type)**

詞次 **sû-chhù (word token)**

eg : "A B A C D B A B C"

4 word types & 9 word tokens

3. 平均詞長 Pêng-kun sû-tn̂g

音節數 / 語詞數

im-chiat sò͘ / gú-sû sò͘

(台灣) 中文： **1.6**

台文：**1.4**

4. 斷詞 Tn̄g-sû Word Segmantation

羅馬字無問題，漢字有

Lô-má-jī bô būn-tôe, hàn-jī ū

(正向&逆向)長詞優先

(sūn-hiòng / tò-thâu) tn̂g-sû iu-sian

(forward / backward) maximal matching

eg1：ùi 聖經看台語語詞變化

(FMM) ùi 聖經 看台 語 語詞 變化

(BMM) ùi 聖經 看 台語 語詞 變化

eg2：(((國民)((大會))(代表))) 人民 行使 職權

有時需要 **heuristics**

Ū-sî-á su-iàu heuristics

eg： 平均詞長

真 正常 用 … **vs** 真正 常用 …

歧義(ambiguity)有的可用統計解決，有的不行

Hâm-hō͘ ê būn-tôe, ū-ê ē-tàng iōng thóng-kè lâi kái-koat, ū-ê bōe-tàng

Boundary word 辭典查不到的詞

sû-tián chhâ bô ê sû

有可能是錯誤，有可能是其它原因

ū khó-lêng sī chhò-gō͘, mā ū khó-lêng sī kî-tha ê gôan-in

專有名詞處理

choan-iú bêng-sû chhú-lí

人名 lâng-miâ

地名 tōe-miâ

機構名稱 ki-kò͘ miâ-chheng

★ 台語還沒有人做 Tâi-gú iah bô-lâng chòe

數字處理 Sò͘-jī chhú-lí

因爲不可能把所有數字放進辭典

In-ūi bô khó-lêng kā só͘-ū ê sò͘-jī lóng khǹg jip̍-khì sû-tián

Regular expression

定量詞處理 Tēng-niû-sû chhú-lí

eg： 一張紙、三尾魚、第五個人

★ 台語還沒有人做 Tâi-gú iah bô-lâng chòe

★ 適合做碩士論文 sek-ha̍p chòe sek-sū lūn-bûn

5．統計公式 Thóng-kè kong-sek

$$MI(AB) = - \log \frac{P(A)\ P(B)}{P(AB)}$$

$$rel(AB) = \frac{n\,(n_{11} \times n_{22} - n_{12} \times n_{21})^2}{n_{1*} \times n_{2*} \times n_{*1} \times n_{*2}}$$

	B	$\sim B$	Σ	AB 表示 A 後壁 sòa 接 B
A	n_{11}	n_{12}	n_{1*}	A~B 表示 A 後壁 m̄是 B
$\sim A$	n_{21}	n_{22}	n_{2*}	~AB 表示 B 頭前 m̄是 A
Σ	n_{*1}	n_{*2}	n	~A~B 表示這兩個音節(詞彙)第一個 m̄是 A 第二個 m̄是 B

音節的 **MI** → 雙音節詞

im-chiat ê MI → siang-im-chiat sû

語詞的 **MI** → 語詞搭配

gú-û ê MI → gú-sû tah-phè **(collocation)**

★ 適合做碩士論文 sek-ha̍p chòe sek-sū lūn-bûn

6．詞頻統計 Sû-pîn thóng-kè

計算語言學的基礎

kè-sǹg gú-giân-ha̍k ê ki-chó͘

(語料量：漢羅 5M+/全羅 3M+ 音節)

(BMM algorithms)

(有改進空間)

7. Zipf's law

計量語言學基礎

kè-liông gú-giân-ha̍k ki-chhó͘

$$f \propto \frac{1}{r}$$

f : freq　　r : rank

單位：語詞**(word)**

Mandelbrot's formula

$$f = P(r+\rho)^{-B}$$ **, or**

$$\log f = \log P - B\log(r+\rho)$$

每個語言的 P, B, ρ 不同

Múi chit-khóan gú-giân ê P, B, ρ bô kâng

英文：**P=$10^{5.4}$, B=1.15, ρ =100**

★ 台語還沒有人做 Tâi-gú iah bô-lâng chòe

$$m \propto \sqrt{f} \rightarrow m \propto \frac{1}{\sqrt{r}}$$, m : meaning

不好做 bô-hó chòe

8. 詞類 sû-lūi

9. 詞義

10. 語詞對譯