2012

漢語方言語音資料庫自動擴增補完方法

林居正[1], 王昱鈞[2], 蔡宗翰[3]

[1]中央研究院資訊科學研究所

[2]國立台灣大學資訊工程學研究所

[3]元智大學資訊工程系

大綱

- 動機
- 相關文獻
- 資料
- 方法
- 實驗
- 結論

2012

DADHIC

動機

- 數以百計的漢語族語言, 其用語人無法互相通話
- 這些語言皆稱為**漢語方言**

(Tong, 2004)

動機

- 現象
  - 除了普通話以外, 其餘漢語方言資源皆極為匱乏
- 比較: Wikipedia條目數
  - 西班牙語: 846,523
  - 加泰隆尼亞語: 357,960

語言	Wikipedia條目數
英語	3,808,991
普通話	386,850
粵語	16,943
閩南語	9,759
贛語	6,022
吳語	2,872
客語	2,264

動機

- 數位資源: corpora, treebank, language models, POS tags, ...
- 一項很重要的數位資源: 語音資料庫
  - 對各項語音處理應用都極為重要
  - 需要母語者, 但目前方言母語者逐漸凋零
- 在不做田野調查 (可能極為昂貴) 的情況下, 我們可以填補未知的語音嗎?
- 文化遺產保存

相關文獻

- 漢語方言TTS
  - Lin and Chen, 1999
- 資源匱乏語言各項應用
  - Snyder et al., 2007
- 缺漏值的資料增補
  - Zhu, 2005; Lin, 2005
- 歷時音韻學
  - Bouchard-Côté et al., 2007

2012

DADHIC

資料

- DOC dataset
  - 1960年代在 UC Berkeley 進行的 Project DOC (Dictionary On Computer) 收集的資料
- 目的是幫助歷史語言學研究
- 整理過後得到3,403個讀音
  - 區分多音字: 如 “正” (zheng1, zheng4)

資料

- 21個方言
  - 涵蓋所有大方言群: 官話, 吳語, 粵語, 湘語, 客語, 贛語, 閩語
  - 以 IPA 音標轉寫; 重新以 8 個音韻特徵轉寫
- 每個漢字音的韻書類別特徵
  - 韻鏡, 廣韻

資料-
IPA轉寫的音韻特徵

說(廈)[suat˩]

調類	入
調值	2
聲母	s
介音	u
元音	a
雙元音後半	∅
鼻化	否
韻尾	t

資料-
韻書類別特徵

聲母	書
調	入
呼	合
韻母	薛
攝	山
等	三等

說

方法

- 問題定義
- 資料增補
- 模型應具備性質
- 模型定義
- 推論

2012

DADHIC

問題定義

- 填補語音資料庫內之遺缺語音
  - 視為一資料增補問題
- 輸入
  - 每個漢字音的韻書特徵
  - 每個漢字音在一些方言中實際語音
- 輸出
  - 每個漢字音於每個方言的實際語音

資料增補 (Tanner and Wong, 1987)

- y 為觀察到不完整的資料
- θ 為待增補之遺缺資料

- 若機率分布 $P(\theta \mid y)$ 不好計算, 則

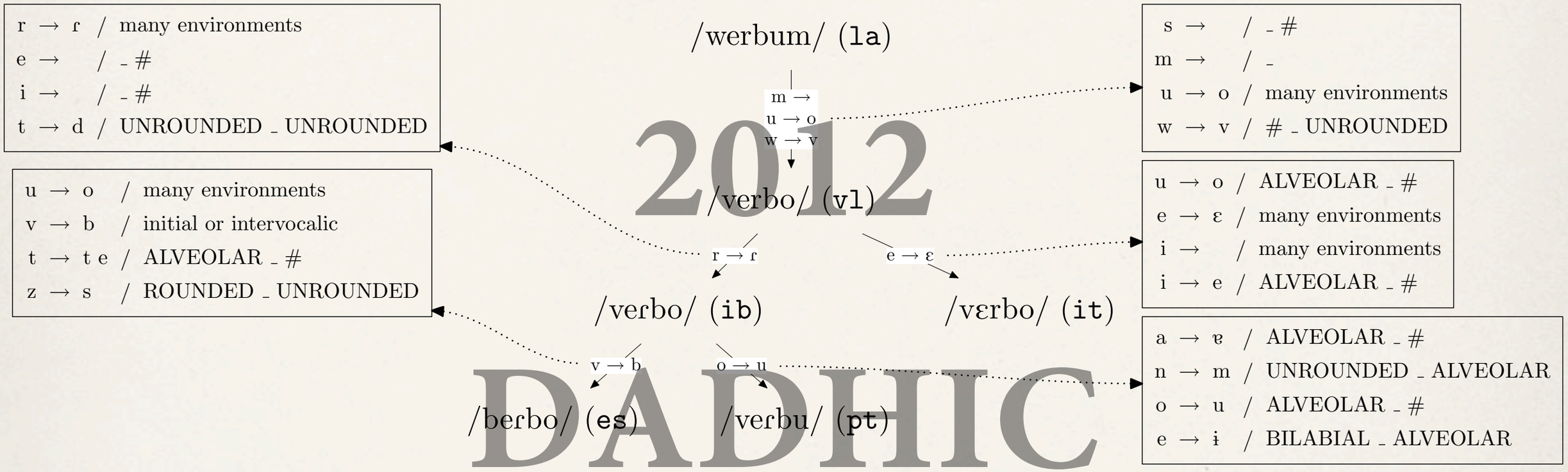
  - 試著舉出隱藏隨機變數 z 使 $P(z \mid y,\theta)$ 與 $P(\theta \mid y,z)$ 易處理
  - $P(z, \theta \mid y)$ 可透過馬可夫鏈蒙地卡羅 (MCMC) 法得到, 而
  - $P(\theta \mid y) = \int P(z, \theta \mid y)\, dz$

模型應具備性質

- 就我們的資料而言
  - 韻書類別特徵總是存在
    - 視為監督式學習方式的特徵值
  - 現代方言音韻特徵可能遺缺
    - 以資料填補觀點，透過導入隱藏變數得到

字	韻	北	廣	潮	…
肝	見寒山平開一	kan	kon	kua~	…
圖	定模遇平合一	thu	thou	?	…
…	……	…	?	…	…
…	……	…	…	…	…
…	……	…	…	…	…
…	……	…	…	?	…
…	……	…	…	…	…
…	……	…	…	…	?
…	……	…	…	…	…

模型應具備性質

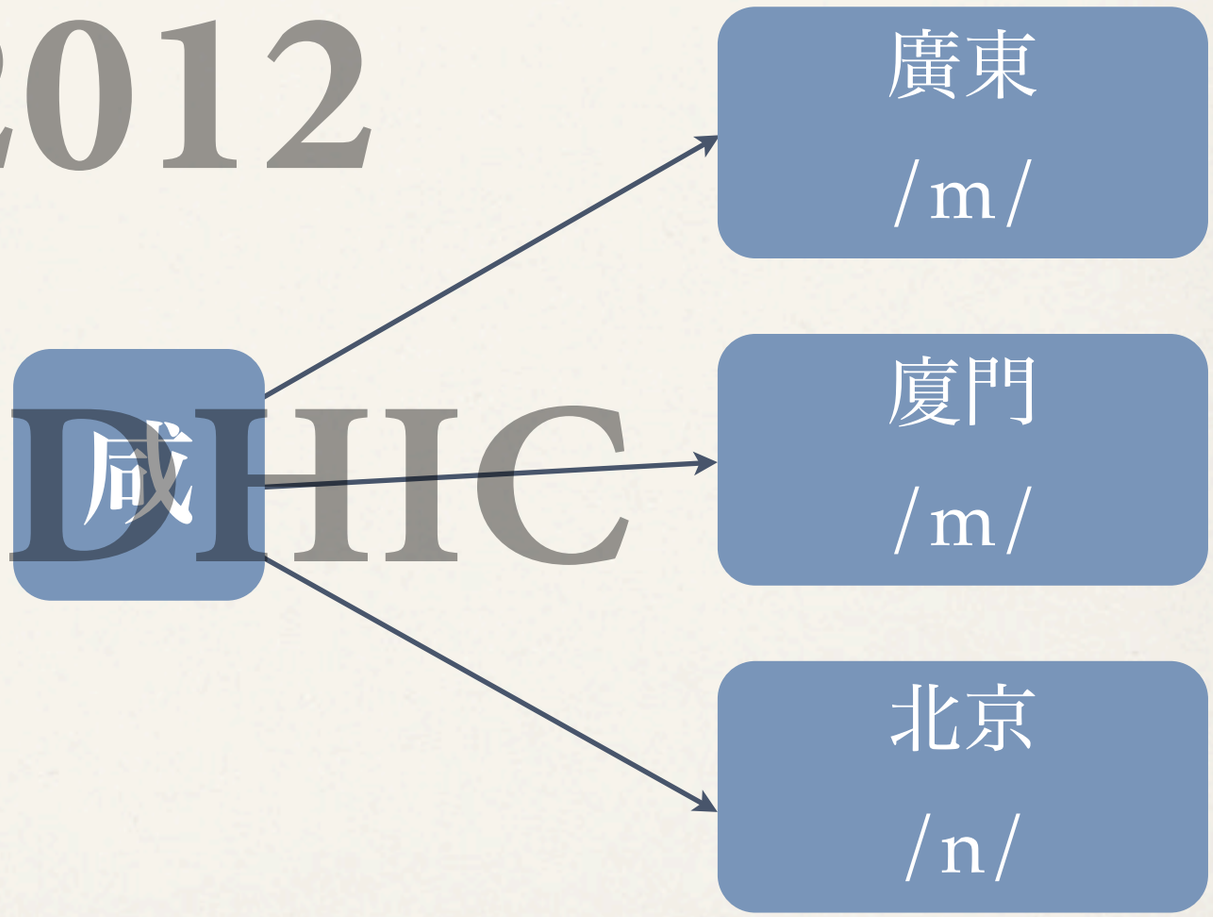


- Bouchard-Côté 等提出有親緣關係的語言模型; 但他們的模型要求
  - 先行構擬發生樹
  - 祖語的語音轉寫
- 這兩項要求使他們的模型不適用於我們的資料

模型應具備性質

2012

- 韻書資料與現代方言音韻特徵對應

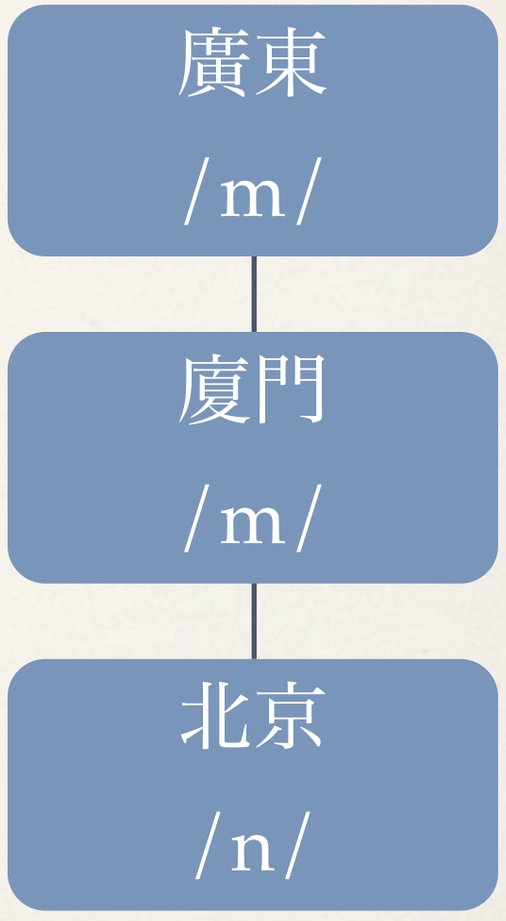


模型應具備性質

2012

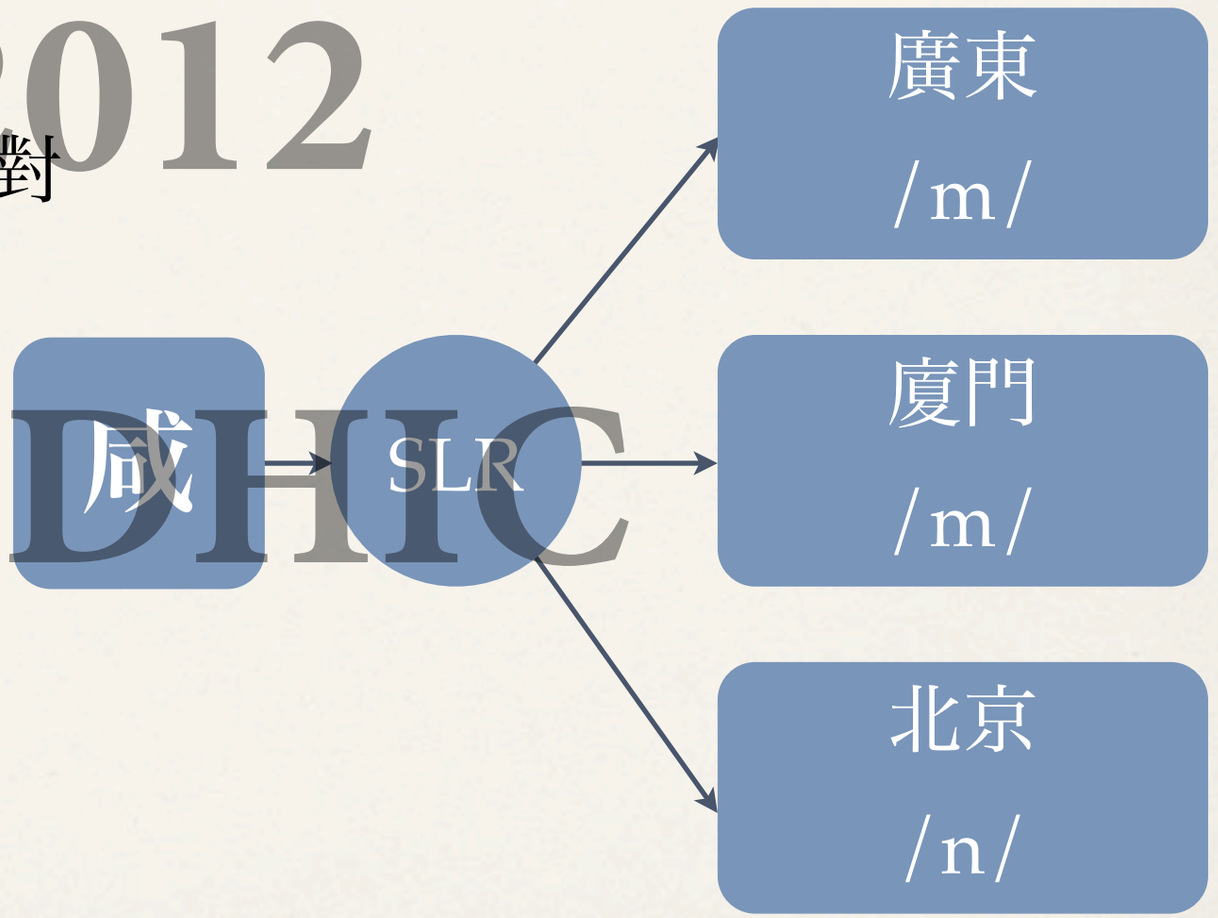
DADHIC

- 方言之間的音韻特徵對應



模型應具備性質

- 韻書資料與現代方言音韻特徵對應
- 方言之間的音韻特徵對應
- 以隱藏的 *superlingual rhymes* (SLRs) 模型表示



模型定義

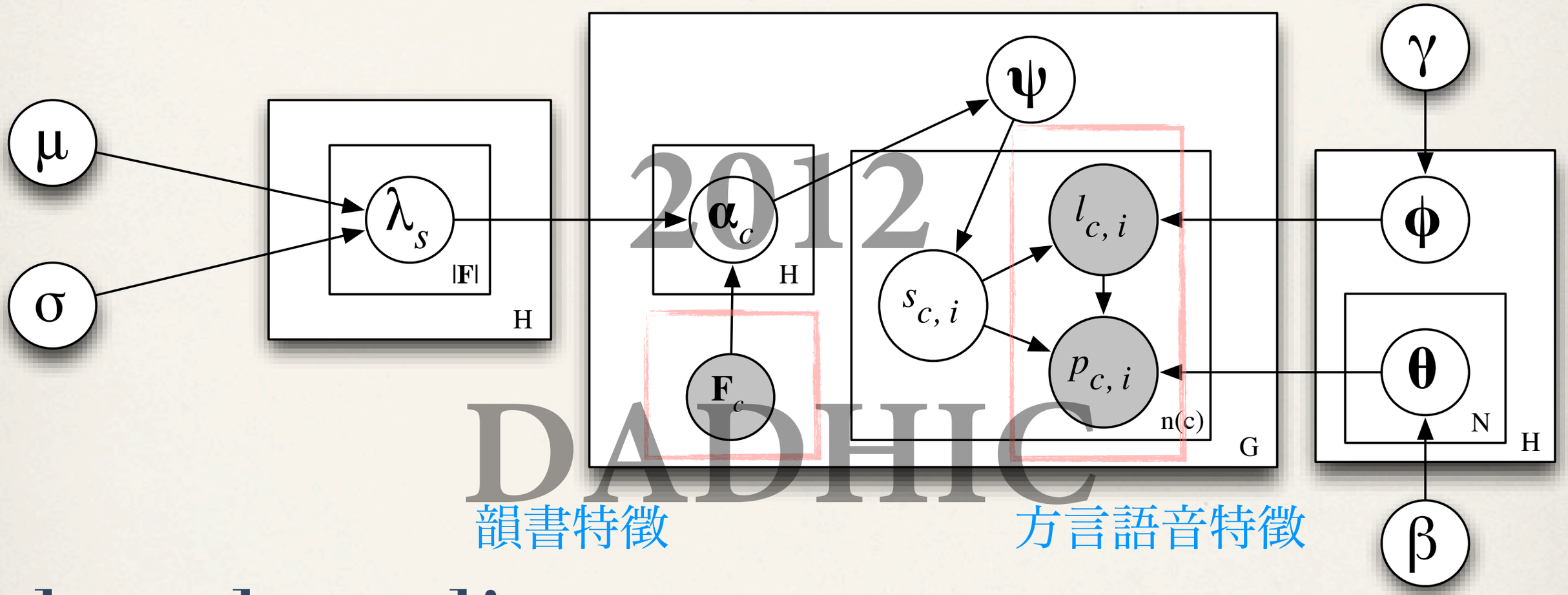
2012

- 生成模型 (generative model)

DADHIC

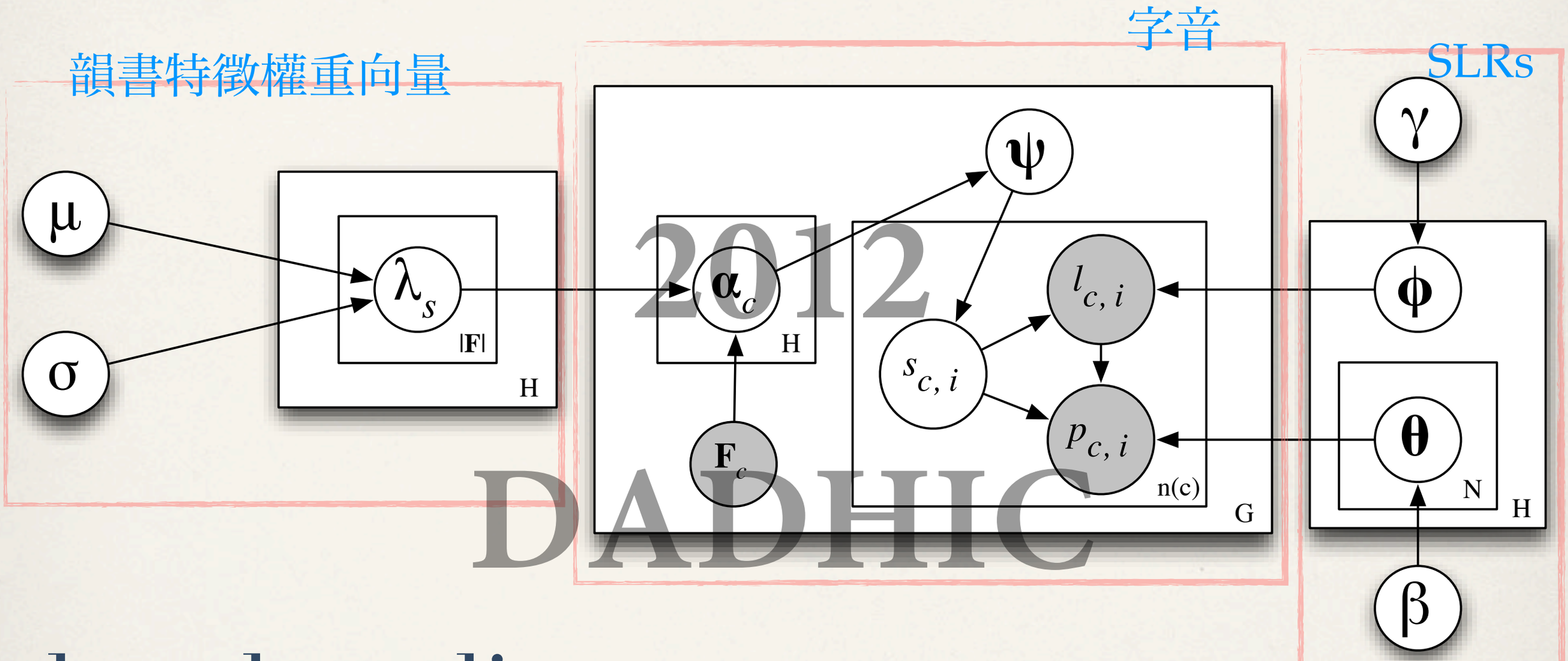
- 每個音韻特徵皆對應到某隱藏 *superlingual rhyme (SLR)*

- 字音的韻書特徵限制該字音會有哪些 SLR

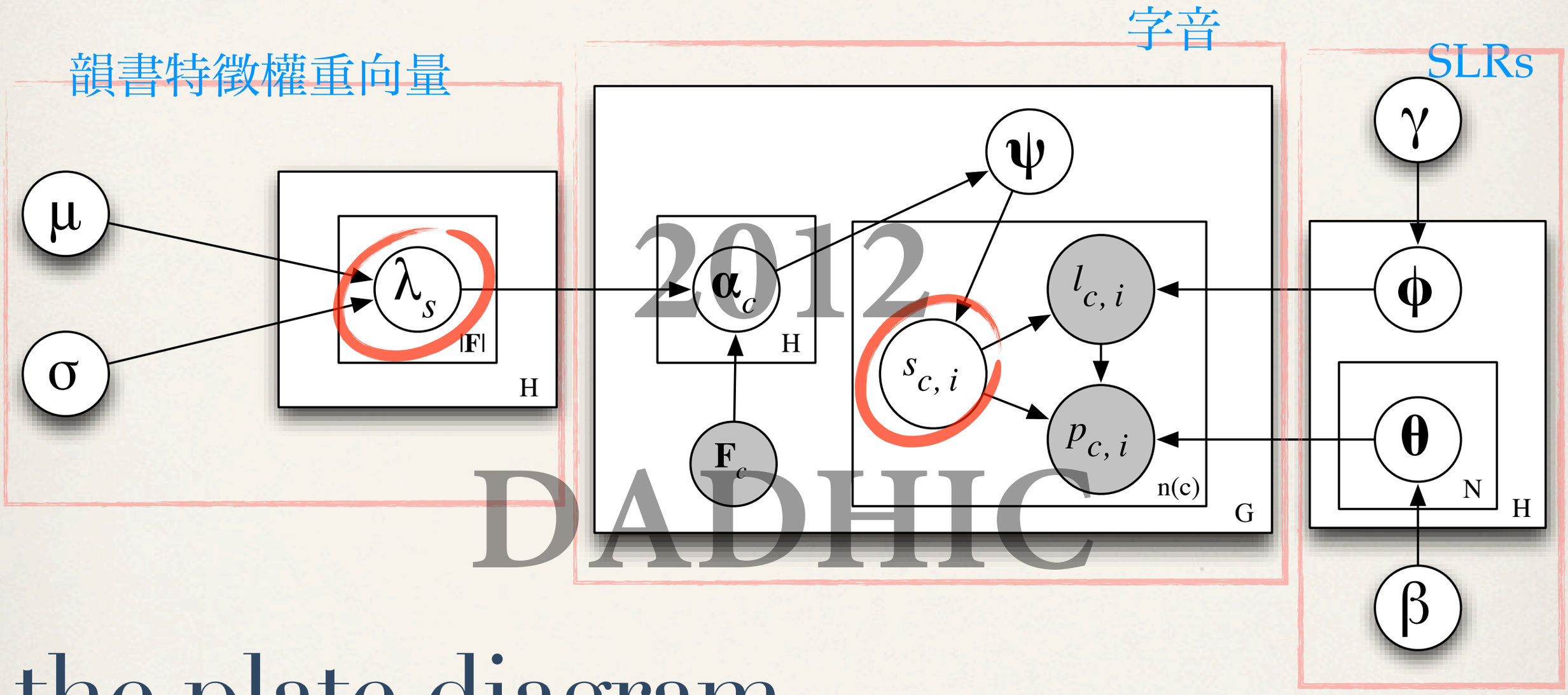


韻書特徴　　方言語音特徴

the plate diagram



the plate diagram



the plate diagram

推論

- 可分兩部分
  - 每個觀察值 $p_{c,i}$ 的 SLR $s_{c,i}$
  - 每個 SLR s 的 λ_s
    - s 對應到哪些韻書特徵
- 大致流程
  - 從 $P(s_{c,1} \ldots s_{c,n(c)})$ 用馬可夫鏈蒙地卡羅法 (MCMC) 取樣; 固定 λ
  - 固定 $s_{c,1} \ldots s_{c,n(c)}$; 最大化 L(λ's; everything else)

實驗

2012

- 實驗設定
- 方言資料對標準分類器的影響
- 鄰近方言的影響
- 資料填補的效果

實驗設定

- 評估指標: 整體音韻特徵準確率 (**Overall Pronunciation Feature Accuracy**, ***OPFA***)
- 評估方法
  - 目標語言以外的方言隨機移除音韻特徵至固定缺漏數
  - 用我們提出的方法或baseline演算法填補空缺值
- 以2:1比例分為訓練資料集與測試資料集
- SVM 分類器
  - RBF kernel
  - parameters: c = 512, gamma = 2^{-7}
- superlingual rhyme 數目設為 200

方言資料對標準分類器的影響

- 僅有韻書 (R)
  - 僅包含韻書特徵: “聲母”, “韻”, “攝”, “聲調”, “呼” 以及 ”等”
- 韻書與全部方言資料 (R+F)
  - 韻書特徵 + 現有所有方言音韻特徵
- 預測潮州方言

2012

DADHIC

實驗設定	OPFA
R	41.5%
R+F	62.8%

鄰近方言的影響

- 韻書特徵加上關係密切的方言 (R+C)
- 韻書特徵加上關係疏遠的方言 (R+D)
- 官話系統: 濟南, 太原, 北京
- 閩語系統: 廈門, 福州, 建甌
- 目標語言: 西安, 潮州
- 隨機移除音韻特徵至10%缺漏

實驗設定	OPFA
西安 (R+C)	80.5%
西安 (R+D)	67.9%
潮州 (R+C)	51.2%
潮州 (R+D)	42.8%

資料填補的效果

- 與潮州方言關係密切的方言
- 移除至缺漏 20% 音韻特徵
- 分別用我們提出的方法(**Data Augmentation**) 與三組 baseline 填補
  - Logistic Regression
  - Naïve Bayes
  - Random

填補方法	OPFA
Data Augmentation	**54.0%**
Logistic Regression	48.1%
Naïve Bayes	48.3%
Random	46.4%

結論

- 我們提出了一個新的生成模型, 能同時利用中古韻書資料與可能不完整的方言字音資料發掘跨方言的音韻規律 (稱之 superlingual rhymes)

- 我們的實驗揭露
  - 除了韻書資料, 方言資料對於預測另一方言字音有很大幫助
  - 關係密切的方言資料較為有效
  - 透過我們提出的模型填補語音資料, 能提高預測新方言語音的準確率