瀕危語言的數位典藏與加值

楊孟蒨　　靜宜大學

大綱

- 瀕危語言數位典藏
- 線上語料建立
- 語料研究與加值方式

瀕危語言的數位典藏

台灣原住民語言數位典藏

- 台灣原住民語言的重要性
  - 南島語言的發源地
  - 語言之間最分歧
  - 保存最多古音特徵(李 2009)
- 台灣原住民語言典藏工作的重要性
  - 保存台灣原住民語言
  - 鼓勵與支持原住民語言學習

原住民語言典藏計畫相關步驟

- 語料採集
- 語料記音與語料研究
- 語料典藏整理及網路呈現
- 典藏語料的運用與加值
- 典藏語料研究

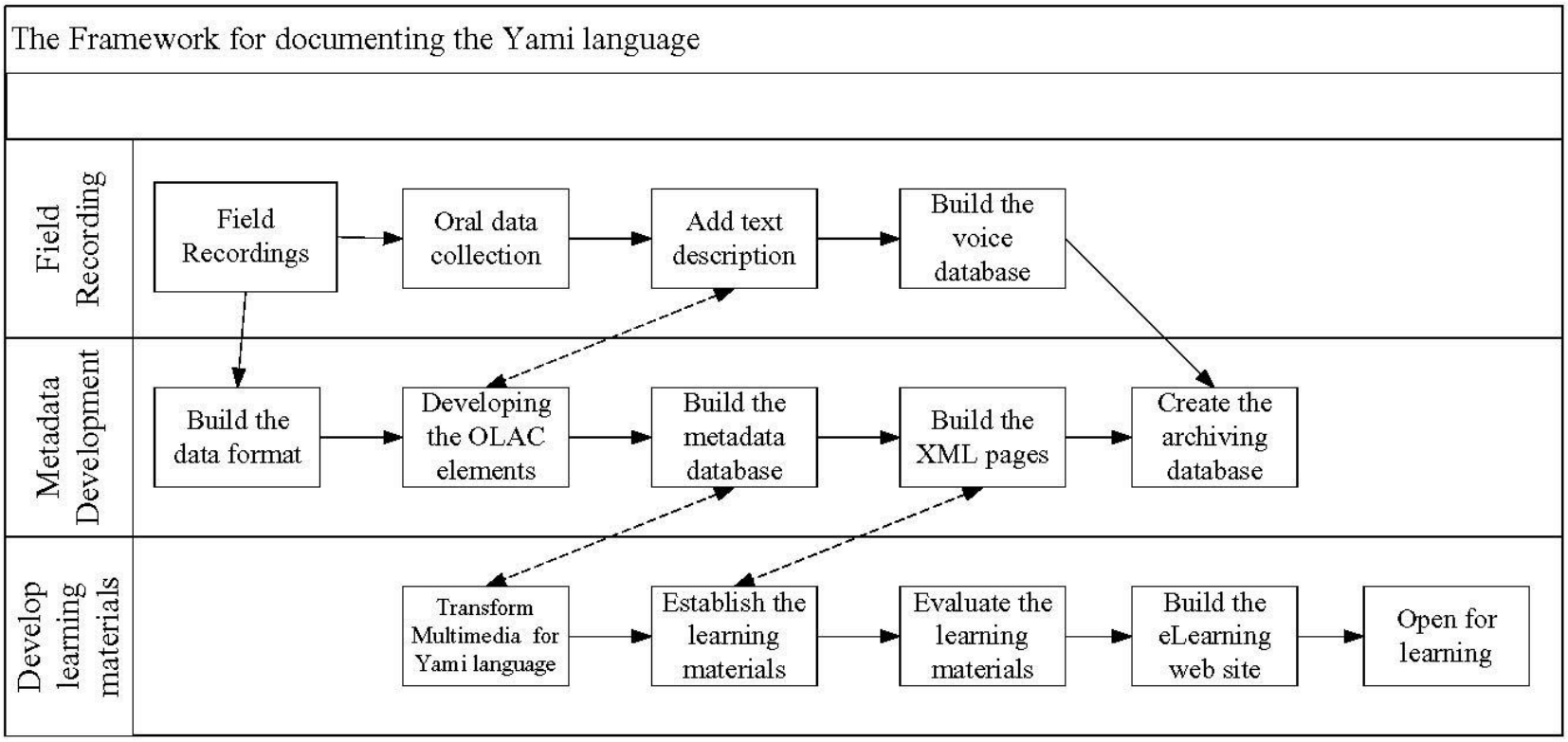
Yami 語言蒐集內容

- Yami narratives
- Daily conversations, business transactions and festivals/ceremonies
- Yami reference grammar
- Trilingual dictionary with 2000 entries
- Multimedia pedagogical materials
- Folk songs

達悟語數位典藏計畫重要步驟

1. 田野調查與錄音
2. 建立數位化資料
3. 建立多媒體資料
4. 建立線上數位典藏語料庫
5. 建立線上學習系統
6. 建立線上辭典
7. 建立線上語意網

數位典藏架構圖


The Framework for documenting the Yami language
Field Recording
Field Recordings
Oral data collection
Add text description
Build the voice database
Metadata Development
Build the data format
Developing the OLAC elements
Build the metadata database
Build the XML pages
Create the archiving database
Develop learning materials
Transform Multimedia for Yami language
Establish the learning materials
Evaluate the learning materials
Build the eLearning web site
Open for learning

已建置瀕危語言語料庫

- 達悟語線上典藏網
- 達悟語線上學習網
- 達悟語線上辭典
- 達悟語線上語意網

達悟語線上典藏網

Yami language archive

http://yamiproject.cs.pu.edu.tw/yami

這是最早以田野調查的語料所建置的語料及詮釋資料查詢系統

達悟語線上學習網

Yami e-Learning

http://yamiproject.cs.pu.edu.tw/elearn

- 這是一套分為三種等級的達悟語數位學習系統

達悟語線上詞典

- 由達悟族語專家、南島語言學者、和資訊顧問共同合作，為具備中文語言背景、欲學習達悟語言者所編纂之工具書。

http://yamibow.cs.pu.edu.tw/

達悟語線上語意網

- 這是一套展示語意分析及語料加值方式的研究成果展示系統
- http://yamionto.cs.pu.edu.tw/

瀕危語言數位加值以達悟語為例

瀕危語言的語料運用研究:語意加值

- 運用瀕危語言語料來研究瀕危語言的自然語言處理
- 研究動機: 設計瀕危語言的語意網 (semantics web)
- 所遭遇到的困難:
  1. 很少計算語言的演算法或模式針對於原住民語言來研究
  2. 與中文知識庫比較，相對於原住民語料的網路資源可以說相當少

研究加值方式: 建立知識本體

- 運用本體論來建構知識，並且彌補語料的不足
- 透過不斷的分析與實驗，並且使用語言學的知識來改善演算法可能沒考慮的問題

建構原住民語知識本體(以達悟語為例)

知識本體(ontology)

- Ontology一詞是由哲學發展出來的，是在說明在特定實體會依據共同的性質來分類，表指把實體概念化。
- 建立Ontology是為了人與軟體間能有效的分享資訊結構，透過結構可以重覆知識，並且透過一再學習找出一條創新的規則。

（例：基本(已知)規則 － 1+1=2

套用(創新)規則 － 1蘋果+1蘋果=2蘋果）

知識本體組成元素(1/1)

- 概念（Concept、Object、Class）
  - 表所有字彙所組合而成的集合。

[例：Person → adult、juvenile]

- 屬性（Attribute、Property、Slot、Role）
  - 描述字彙的特性或特徵。

[例：rarakeh 有重疊（hasReduplicate）及有相反詞（hasConceptReverse）特徵]

知識本體組成元素(1/2)

- 關係（Relation）
  - 字彙彼此的關係。

[例：rarakeh hasReduplicate ra
或 rarakeh “老人” hasConceptReverse kanakan “孩童”]

- 實例（Instance）
  - 實例可以表達上層（concept）的意義。

[例：adult - rarakeh “老人” juvenile - kanakan “孩童”]

從魚類詞彙庫到魚類知識本體

- **將實體概念化 = 達悟語魚類名稱實體概念化**
- Toolbox → Lexique Pro → Protégé

作法五步驟（Protégé）

1） 決定所要做的主題領域知識本體

達悟語魚類名稱

2) 利用所現有的資源

“雅美(達悟)族的海洋生物”

台灣魚類資料庫（http://fishdb.sinica.edu.tw/）

3) 列出所要設的本體名稱/項目

可食用,不可食用;人(男人,女人,老人, 懷孕婦人);

魚（男人魚, 女人魚,老人魚）

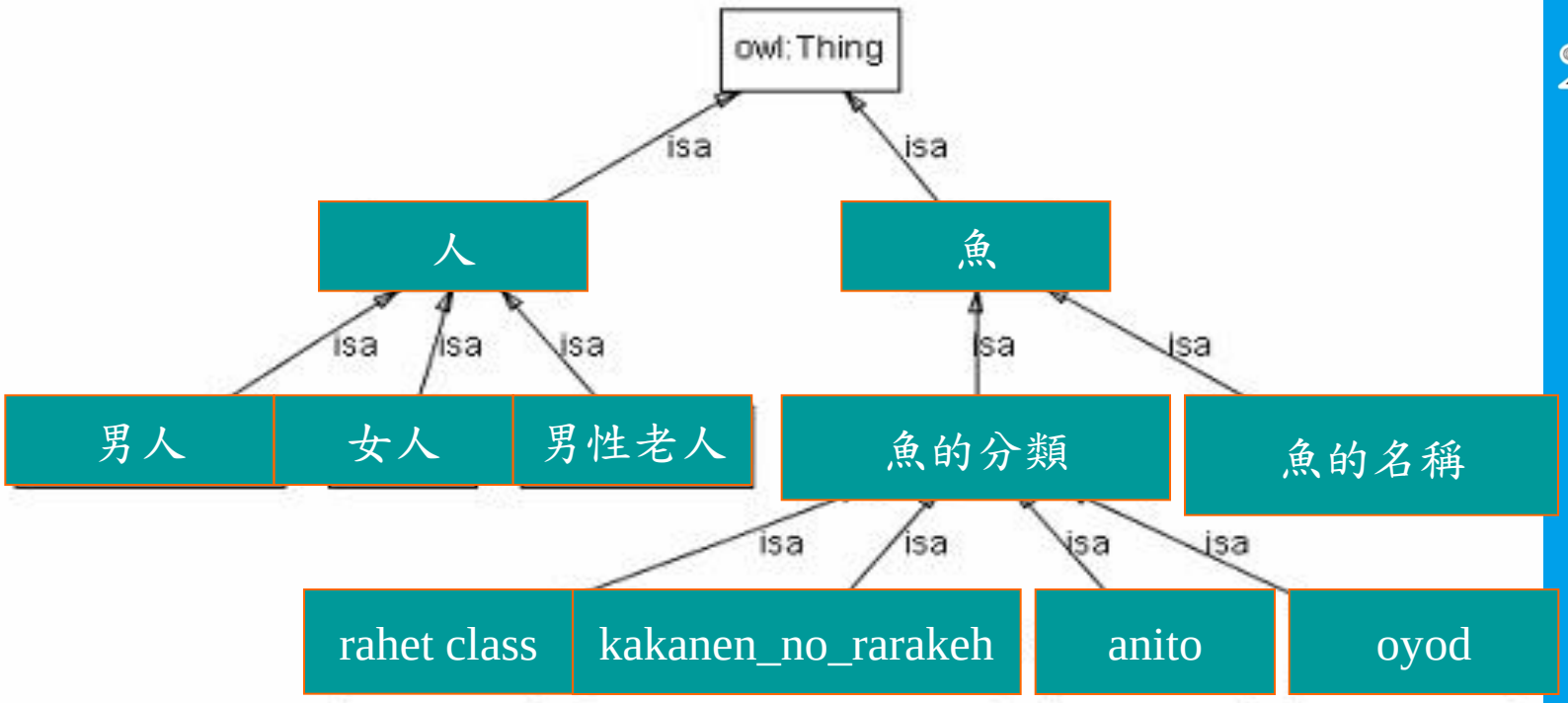
4) 定義類別及類別階級及定義類別屬性

可食/不可食、中文科名/中文俗名…

5) 建立個體

建立輸入魚類名稱

輸出連結關係

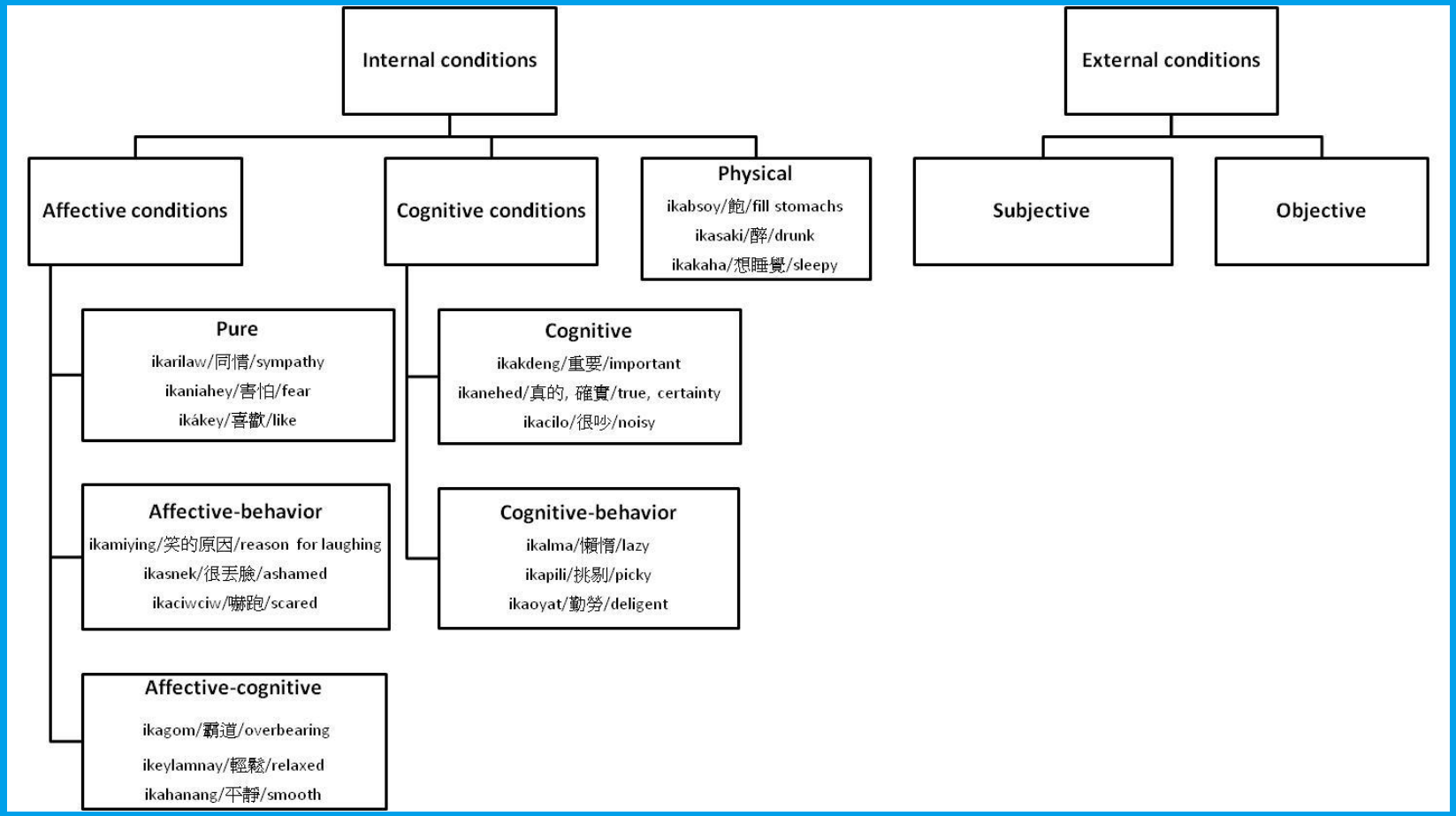

owl:Thing
isa
isa
人
魚
isa
isa
isa
isa
isa
男人
女人
男性老人
魚的分類
魚的名稱
isa
isa
isa
isa
rahet class
kakanen_no_rarakeh
anito
oyod

研究成果: 使用知識本體來分析達悟語的情緒語詞

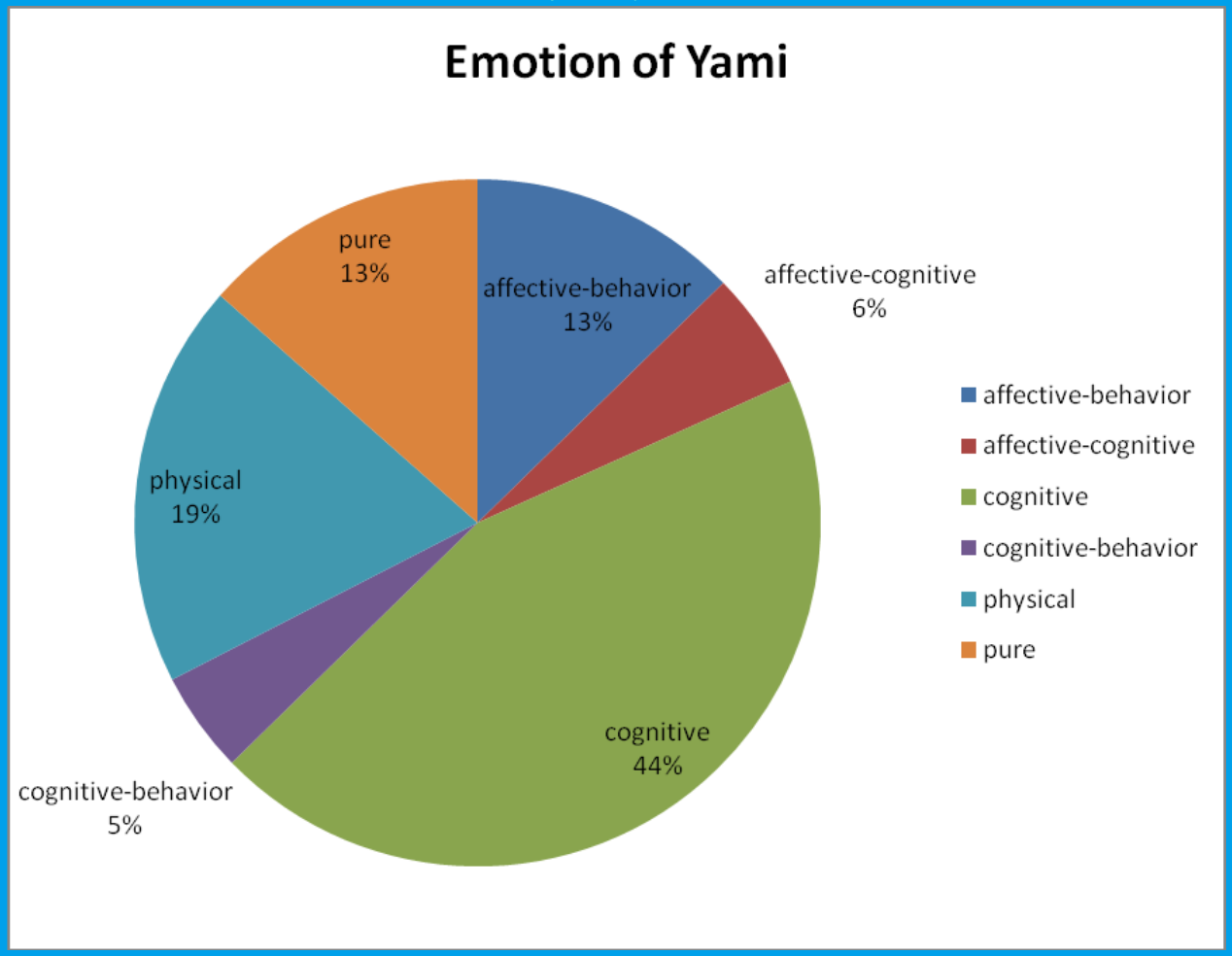
Study of emotions

- **Most important and least important emotions**
- The previous study by Church et. al.'s(1988) about the emotional phrases in Filipino.
- Can the important emotions and the minimally lexicalized emotion domains in Yami generally match Church et al.'s (1988) findings in Filipino data?

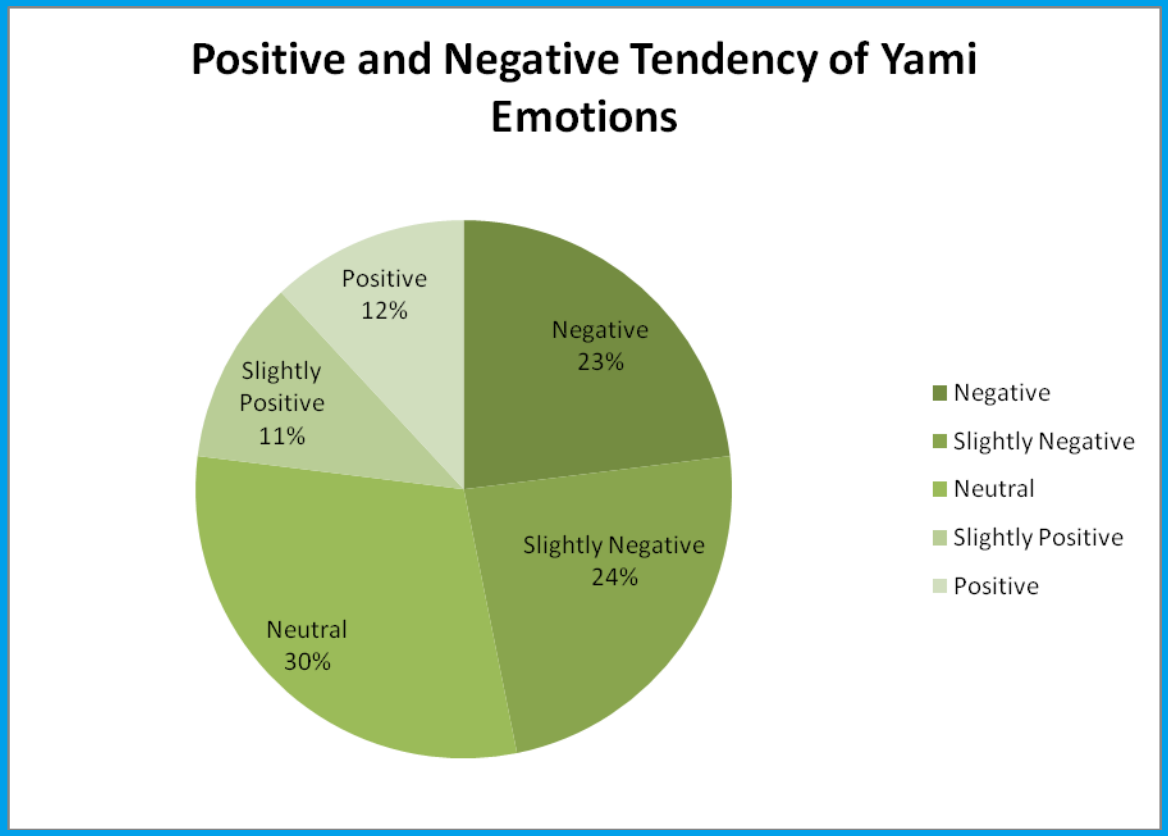
Hierarchy of Emotions



由語料與文獻所得達悟語情緒語詞的分類



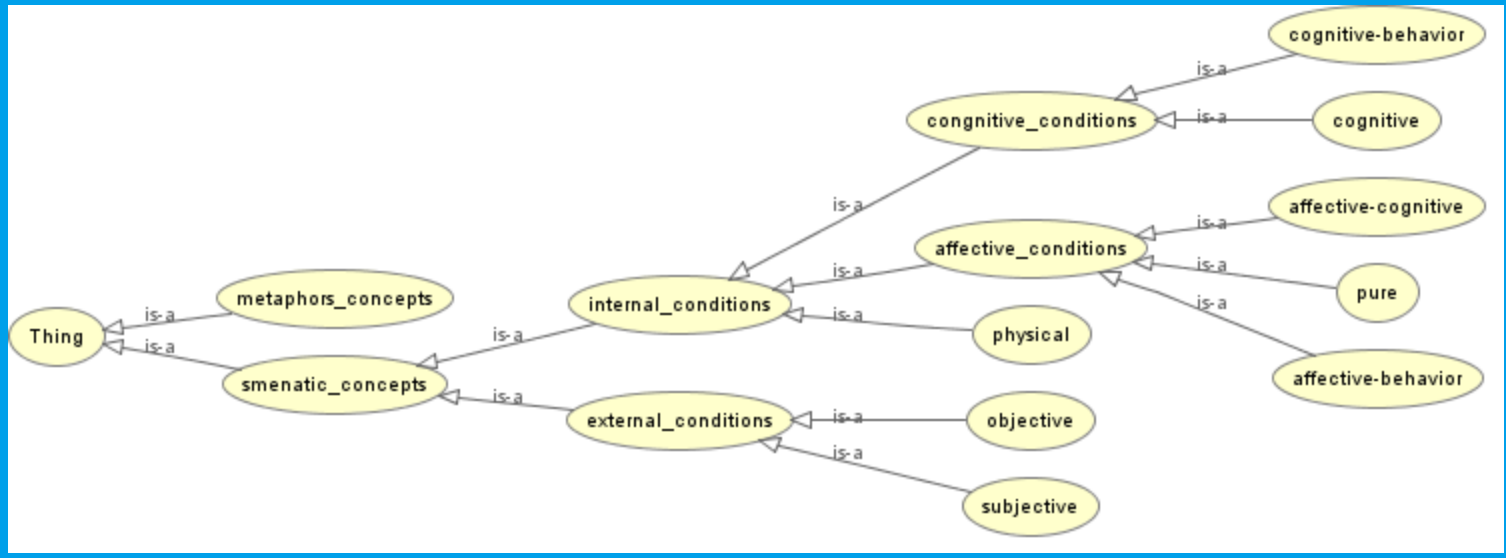
Positive and Negative tendency



Ontology for Yami Emotional Phrases

- 運用語料及文獻分類為基礎
- 設計二個知識本體來分析及歸納情緒語意
  - Forward semantics ontology: 以原先分類架構來設計
  - Backward semantics ontology: 使用隱喻及相關結構反轉回原來語意來分析

Ontological Representation for Emotions



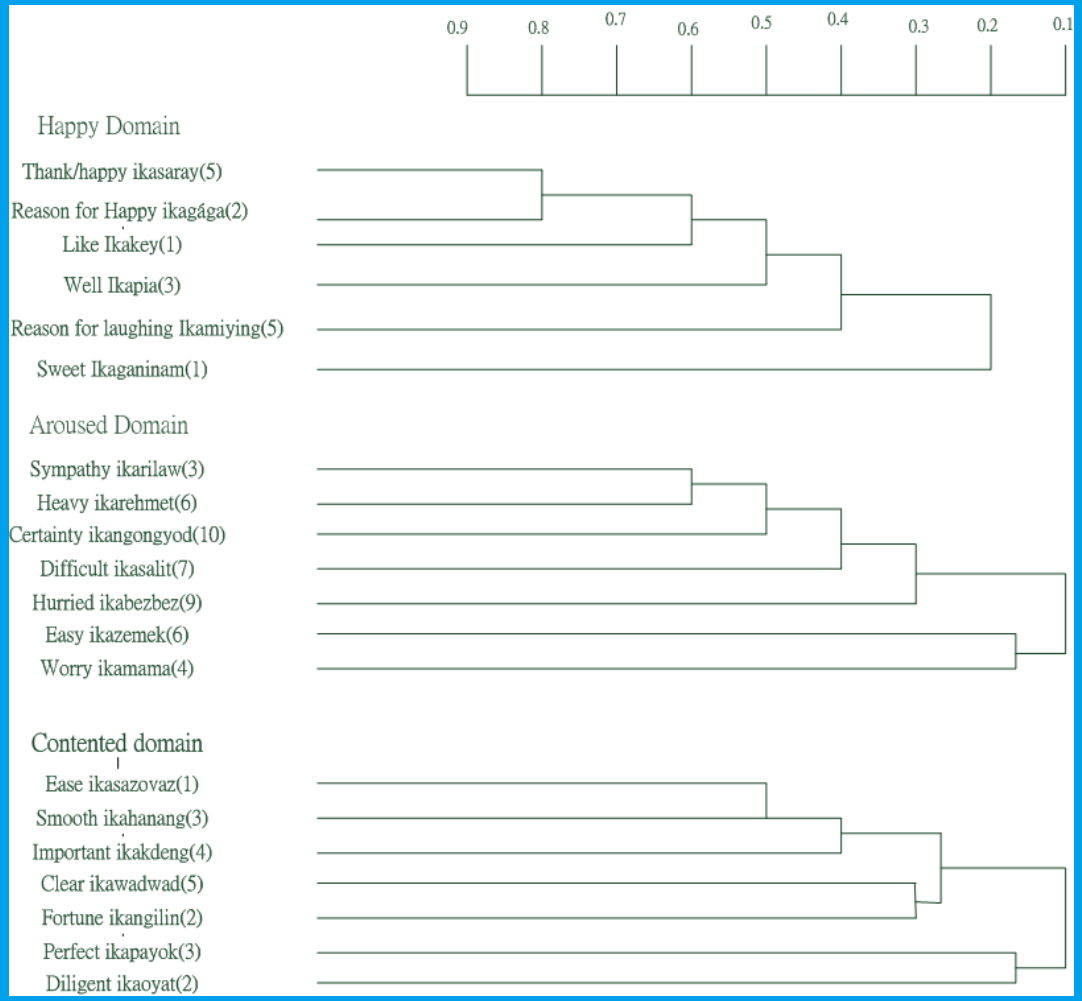
Processing and Evaluating

- Derive an approach for calculating the semantic domains for the Yami phrases using the ontological representations
- The weighted factors are assigned to each Yami phrases based on its the upper-level semantics linked to the English meanings in the metaphorical ontological representation and the analyzed results
- Clusters for these Yami phrases are calculated following the Church's domain

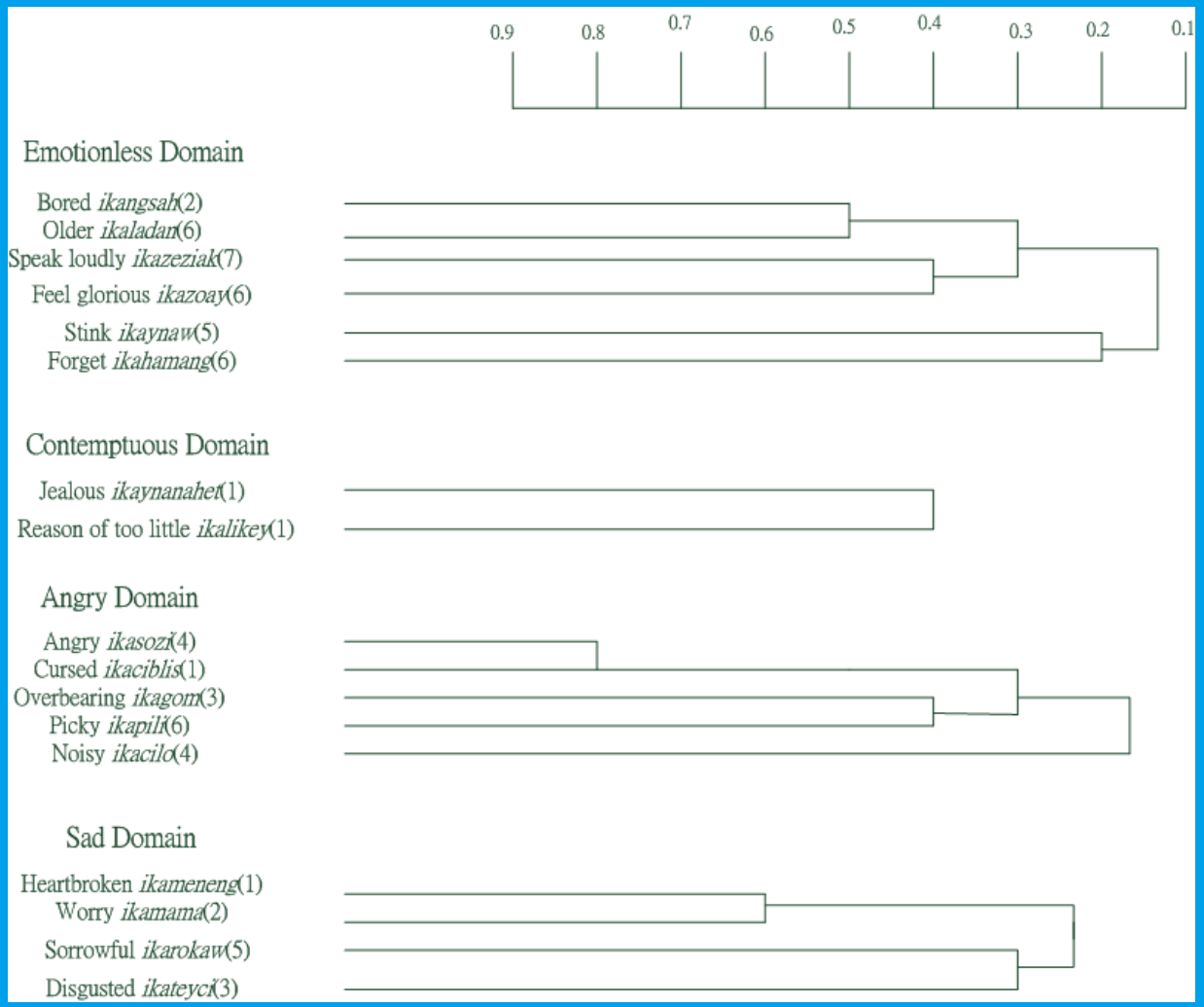
Concepts of Clusters

Domain Concepts	Number of ika-emotional phrases	degree of similarity
Happy Domain	20	[0.8—0.2]
Aroused Domain	45	[0.6---0.2]
Contented Domain	21	[0.8---0.2]
Emotionless Domain	34	[0.5---0.2]
Contemptuous Domain	2	[0.4---0.2]
Angry Domain	19	[0.8---0.2]
Guilty Domain	2	[0.3---0.2]
Sad Domain	11	[0.8---0.2]
Tired Domain	5	[0.6---0.2]
Quiet/Shy Domain	3	[0.4---0.2]
Anxious Domain	55	[0.7---0.2]
Aspiring Domain	2	[0.4---0.2]

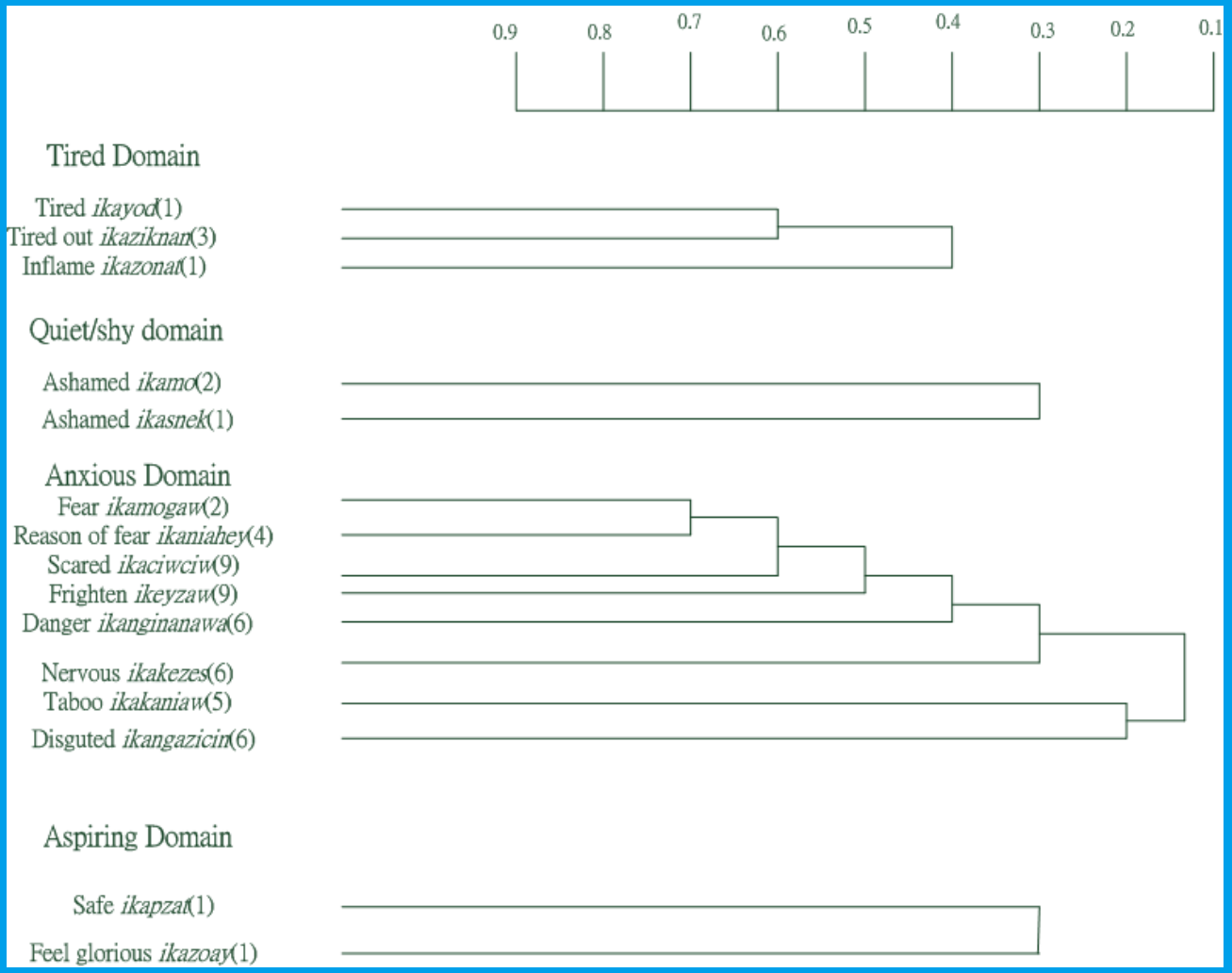
Classification Results(1)



Classification Results(2)



Classification Results(3)



結語

- 我們討論由數位典藏的語料來進行進一步研究的可能性
- 我們探討可能的瀕危語言計算及自然語言處理方式
- 使用知識本體來做瀕危語言的自然處理
- 運用知識本體來測試與討論達悟語的 emotional phrases

Ayoy!

謝謝