

コードのように台湾語を解析

Rubyによる白話字ローマ字の3段階解析

鄧慕凡 (Mu-Fan Teng)

RubyWorld Conference 2025

島根県立産業交流会館 「くにびきメッセ」 Nov. 7, 2025

自己紹介

鄧慕凡 (Mu-Fan Teng)

- 日本では竜堂 終と呼ばれています
- 5xRuby CO., LTD 創業者
- 台湾のRuby伝道師
- RubyConf Taiwan Chief Organizer
- 三度目のRubyWorld登壇 (2015, 2023, 2025)



RubyCityMATSUE 縁結びの地との10年の物語



5xRubyについて

「愛する技術で愛される製品を創る」

- **創業:** 2014年（台北）
- **専門:** Ruby/Railsを中心としたソフトウェア開発
- **実績:** スタートアップ向けシステム開発を中心に、政府機関との協業案件も手がける



5xRubyの事業



1. 委託開発サービス

- 台湾最大級のRuby開発会社（2014年創業）
- クラウド・オンプレミス両対応のインフラ運用
- 日本・米国・シンガポールを含む国際展開
- スタートアップから上場企業まで長期パートナーシップ
- <https://5xruby.com/en>



2. SOSI製品

- セキュアリモートアクセス管理システム
- 踏み台サーバー（Bastion）機能
- ブラウザベース VDI ソリューション
- <https://www.sosi.com.tw>

アジェンダ

本日の内容

1. 無人入札の物語

- なぜ誰も手を出さなかったのか？

2. 台羅（POJ）とは？

- 台湾語のローマ字表記

3. 分詞アライメント処理の実装

- GSUB による実装

4. Parserとの出会い

- Parslet による再実装

5. プロジェクトの成果

- Ruby の強みと実績

6. まとめ

- 結論

スライド資料

<https://rwc2025.ryudo.tw> (日本語)

<https://rwc2025.ryudo.tw/en> (English)



無人入札の物語

なぜ誰も手を出さなかったのか？

台湾政府案件の特殊性



技術の制約

- Microsoft製品への依存
- .NET/MS-SQL/Windows Server
- Ruby/Railsは落選しがち



プロセスの問題

- RFP（要求仕様書）の不備
- 担当者の専門知識不足
- 実務との乖離



隠れたコスト

- 膨大な文書作成業務
- セキュリティ監査・脆弱性診断
- 現地対応が必須の運用作業

8連敗からの学び



落選の理由（技術以外）

- Microsoft製品前提の仕様
- 既存システム、インフラとの「互換性要求」
- 評価基準の不透明さ
- 価格競争ではなく、技術スタックの制約



9回目：驚きの展開

- 競合：ゼロ
- 「なぜ誰も入札しないのか？」
- 担当者も困惑：「本当に大丈夫ですか？」
- 一体何が起こったのか？

落札後の真相

「分詞（文字分割）が煩雑すぎて
誰も手を出さない」

台羅 (POJ) とは？

日本語との類似性から理解する

台羅（台湾閩南語ローマ字）とは？

前後文脈（漢字）

去日本食壽司

前後文脈（POJ）

khì Jit-pún tsiāh sú-sih

香港、澳門…、臺灣俗日本

Hiong-káng, Ò-mâng...Tâi-uân kah Jit-pún

的時，日本義工共臺灣人

ê sî, Jit-pún gī-kang kā Tâi-uân-lâng

台湾語のローマ字表記

- 正式名称: 臺灣台語羅馬字拼音方案
- 略称: 台羅 (Tâi-lô)
- 制定: 2006年10月、台湾教育部公布
- 地位: 台湾語の公式表記システム

中国語（北京語）ではない

- 台湾語: 閩南語系の言語
- 特徴:
 - 9つの声調
 - 独自の子音・母音体系
 - 鼻音化の表記
- 歴史: 白話字 (POJ) をベースに IPA (国際音声記号) 要素を取り入れて開発

日本語と台湾語の文字システム

日本語のシステム

漢字 → ひらがな

対応関係:

■ 一組の語 → 一組の仮名

例:

生活	→ せいかつ
新幹線	→ しんかんせん
東京駅	→ とうきょうえき

台湾語のシステム

漢字 → POJ

対応関係:

■ 一組の語 → 一組の POJ

例:

繼落	→ suà-lòh
新竹市	→ Sin-tik-tshī
明仔載	→ bîn-á-tsài

共通点: 一組の漢字 ➡ 一組の音標

→ だから「分詞アライメント処理」が必要！

実際の分詞アライメント処理例

入力データ（分詞前）：

- 漢字： 繼落來看新竹市明仔載二十六號的天氣
- POJ： suà-lòh lâi-khuànn Sin-tik-tshī bîn-á-tsài gī-tsap-lak hō ê thinn-khì

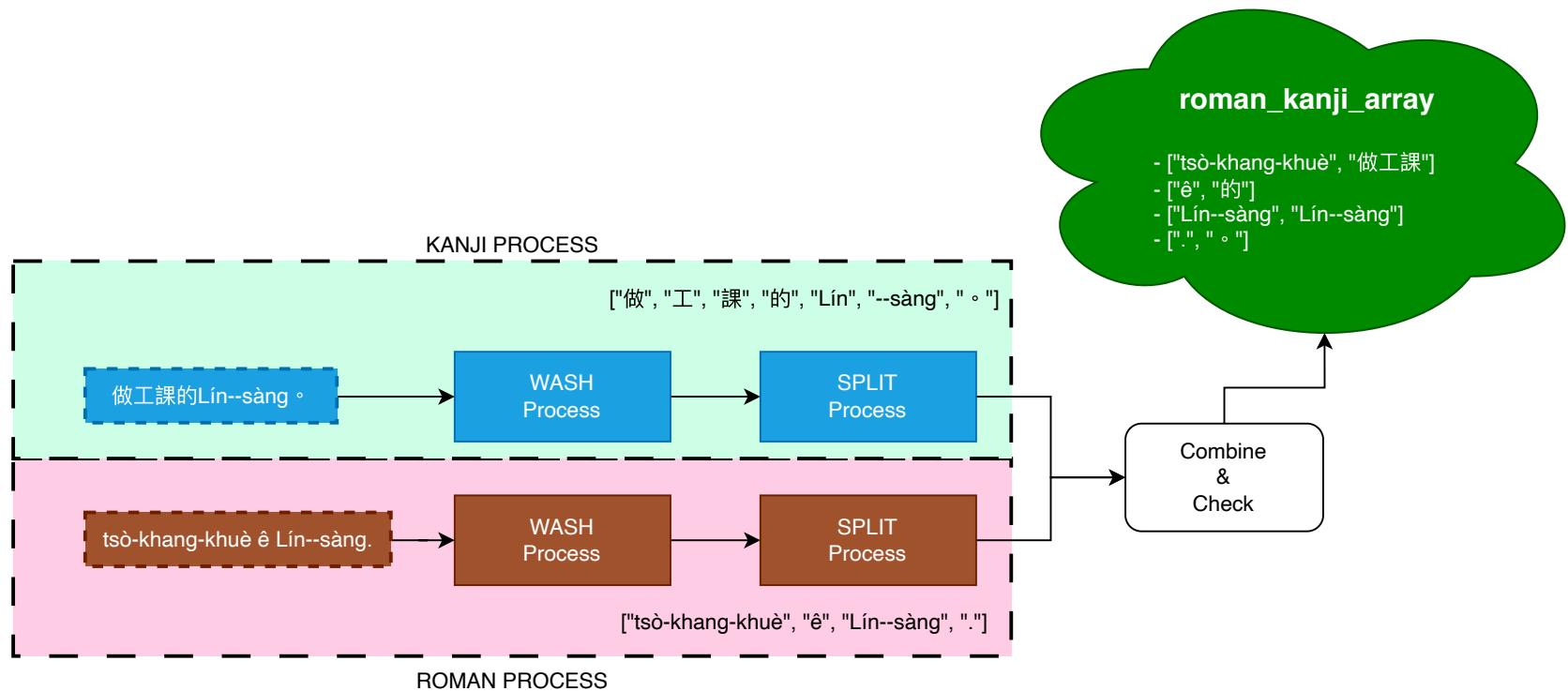
期待される出力（分詞アライメント処理後）：

漢字	POJ
繼落	suà-lòh
來看	lâi-khuànn
新竹市	Sin-tik-tshī
明仔載	bîn-á-tsài
二十六	gī-tsap-lak
號	hō
的	ê
天氣	thinn-khì

分詞アライメント処理の実装

3つのPhaseによる処理フロー

実装の全体フロー：3つのPhase



Phase 1: 正規化 (WASH)

washed_kanji - 漢字側

```
KANJI_GSUB_PATTERNS = {
  /(\w+)(--)(\w+)/ => '\1 \2\3',
  ')((' => ') (',
  /([^\,],)/ => '\1 ,',
  ...
}.freeze
def washed_kanji
  KANJI_GSUB_PATTERNS.reduce(kanji) do |ks, (mt, kp)|
    ks.gsub(mt, kp)
  end
end
```

処理内容:

- 記号の前後にスペース挿入
- ピリオド、カンマ、括弧などを分離
- POJ 文字 (Lín-sàng) も適切に処理

実行例:

入力: 做工課的Lín--sàng。
出力: 做工課的Lín --sàng。

washed_roman - POJ側

```
ROMAN_GSUB_PATTERNS = {
  //"/ => " ",
  /(.)([_+=\:;'"~"\\"[\?!])(.) / => '\1 \2 \3',
  /(^[_+=\:;'"~"\\"[\?!]) / => '\1 ',
  /([\_+=\:;'"~"\\"[\?!])$/ => ' \1',
  /(.)([^\.])/ => '\1 \2',
  /([^\.])(\./) => '\1 \2',
  ...
}
def washed_roman
  ROMAN_GSUB_PATTERNS.reduce(roman) do |rs, (mt, rp)|
    rs.gsub(mt, rp)
  end
end
```

処理内容:

- 65+ パターンで記号を正規化
- ✓ ハイフンは保持 - 音節区切り
- ✓ 二重ハイフン (--) も保持 - 語間停頓

実行例:

入力: tsò-khang-khuè ê Lín--sàng。
出力: tsò-khang-khuè ê Lín--sàng .

Phase 2-1: splitted_kanji - 漢字の分割

実装コード

```
# RXP_SPK - CJK文字と非CJK文字を識別
RXP_SPK = /[\p{Han}\p{Katakana}\p{Hiragana}\p{Hangul}]\u3000-\u303F\uFF00-\uFFEF]|
  [^\p{Han}\p{Katakana}\p{Hiragana}\p{Hangul}]\u3000-\u303F\uFF00-\uFFEF]+/x
ONE_KANJI_WORDS = {
  /(..) (。)/ => '\1\2', /(』) (。)/ => '\1\2', /(--) ([^-])/ => '\1\2' }.freeze
def splitted_kanji
  combine_one_word(
    washed_kanji.scan(RXP_SPK).map do |spka|
      spka.split(/\s/)
    end.flatten.join(' ')
  ).split
end
# combine_one_word - 特殊組合せ処理
def combine_one_word(text)
  ONE_KANJI_WORDS.reduce(text) do |ks, (mt, kp)|
    ks.gsub(mt, kp)
  end
end
```

処理説明

1. RXP_SPK で文字スキャン

- CJK 文字（漢字、ひらがな等）
- 非CJK 文字（POJ、数字等）
- 一文字ずつ or 連続した非CJK文字をまとめて認識

2. combine_one_word で特殊処理

- ONE_KANJI_WORDS パターン適用
- 特定の記号組合せを結合

3. スペースで分割

4. Edge Case の処理:

- Lín--sàng が1つの token として認識される

実行例

入力: 做工課的Lín--sàng 。

出力:

["做", "工", "課", "的", "Lín--sàng", "."]

Phase 2-2: splitted_roman - POJの分割

実装コード

```
def splitted_roman
  washed_roman
    .split(/\s/)
    .compact_blank
end
```

シンプル！わずか3行

処理説明

1. シンプル：スペースで分割

- Phase 1 で記号が既に分離済み
- スペースのみで分割可能

2. 重要な設計:

- ✅ ハイフンでは分割しない
- ✅ 二重ハイフン (--) も保持
- 単語内の音節構造を維持

3. compact_blank で空白要素削除

実行例

入力: tsò-khang-khuè ê Lín--sàng .

出力:

["tsò-khang-khuè", "ê", "Lín--sàng", "."]

音節数:

- tsò-khang-khuè = 3音節
- Lín--sàng = 2音節 (--は音節数に含まれない)

Phase 3: 対齊と検証

```

def roman_kanji_array
  spk = splitted_kanji.dup
  splitted_roman.map do |rword|
    if rword == '--' || (SP_MIRRORS.key?(rword) &&
      #... Edge Case の処理
      [rword, spks]
    end
  end
end

def set_arrays
  rka = roman_kanji_array.transpose
  assign_attributes(
    roman_array: rka[0],
    kanji_array: rka[1]
  )
  self.arrays_balanced = [
    roman_array.size.positive?,
    roman_array.size == kanji_array.size,
    kanji_array.join.size ==
      washed_kanji.delete(' ').size
  ].all?
end

```

処理説明

1. 音節数でマッチング

- ハイフン = 音節区切り
- tsò-khang-khuè (3音節) → 漢字3文字
- Lín--sàng (2音節) → Roman そのまま

2. Edge Case の処理:

- Roman が kanji 側にある場合、そのまま対応
- 二重ハイフン (--) は音節数に含まれない

3. 配列の組み合わせ : transpose で roman/kanji を分離

4. 平衡性検証 (3条件)

- ✓ 配列が空でない
- ✓ roman と kanji の要素数が一致
- ✓ kanji の総文字数が元の文字数と一致

kanji_array: ["做工課", "的", "Lín--sàng", "."]

roman_array: ["tsò-khang-khuè", "ê", "Lín--sàng", "."]

roman_kanji_array [[["tsò-khang-khuè", "做工課"], ["ê", "的"], ["Lín--sàng", "Lín--sàng"], [".", "."]]]

Parserとの出会い

2024年の実装から2025年の気づきへ

金子さんのトークからの気づき

The screenshot shows a presentation slide from the RubyConf 2025 website. The slide has a dark background with a map of Taiwan and the COSCUP logo. At the top left, there's a sidebar with links: SPEAKERS, SCHEDULE, PARTNERS, VENUE, PARTY, and COSCUP. The date '2025/08/09-10' is also listed. The main content area features a large owl icon, the speaker's name 'Yuichiro Kaneko', and a brief bio: 'CRuby committer. The author of https://github.com/ruby/lrama LALR parser generator.' Below the bio are two small circular icons.

TR209 2025/8/9 15:15 ~ 15:45

(EN) Understanding Ruby Grammar Through Conflicts

Have you ever created a parser using an LR parser generator such as Lrama or GNU Bison? If so, you've likely encountered conflicts – perhaps even many times – leaving you feeling lost and unsure of what to do. However, conflicts offer incredibly valuable insights for understanding the grammar of programming languages. In this presentation, we will delve into the structure of Ruby's grammar by intentionally introducing changes to its syntax, triggering conflicts, and then analyzing those conflicts. Through this process, we aim to deepen our understanding of how Ruby's grammar is constructed.

"Understanding Ruby Grammar Through Conflicts"

Parser の3段階処理

1. Lexical Analysis (字句解析)
2. Syntax Analysis (構文解析)
3. Semantic Analysis (意味解析)

💡 **「私がやっていたのは... Parser だったのか！」**

→ **「Parser として実装し直してみよう」**

Conference Driven Development

Parser の方式(ほうしき)で分詞(ぶんし)アライメント処理(しょり)を実装(じっそう)する

Parslet gem との出会い(であり)

Ruby で Parser を書くための DSL ライブラリ

なぜ Parslet ?

- PEG Parser: Parsing Expression Grammar
- Ruby DSL: Ruby の文法で Parser を定義
- 明確な構造: 3-phase 設計を自然に実現

```
# Parslet の基本形
class MyParser < Parslet::Parser
  # Phase 1 & 2: 規則定義
  rule(:word) { match['a-z'].repeat(1) }
  rule(:sentence) { word >> space }

  root(:sentence)
end
```

Parslet の設計思想

Parslet は開発者に **3つの Phase** を意識させる設計：

Phase 1: Lexical Analysis

- rule() で Token 型を定義
- match[], str() で文字パターン

Phase 2: Syntax Analysis

- >>, | で規則を組み合わせ
- 自動的に AST を構築

Phase 3: Semantic Analysis

- Transform クラスで変換
- AST → 最終データ構造

Parslet DSL 基礎語法

基本構文

rule() - 規則の定義

```
rule(:letter) { match['a-zA-Z'] }
rule(:digit) { match['0-9'] }
```

意味: 再利用可能な Parser 規則を定義

match[] - 文字クラス

```
match['a-z']          # a-z
match['a-zA-Z0-9']    # 英数字
match['\u0300-\u036F'] # 声調記号
```

意味: Regular Expression の [...] と同じ

str() - 文字列マッチ

```
str('-')      # ハイフン
str('--)     # 二重ハイフン
str(' - ')   # スペース-ハイフン-スペース
```

意味: 文字列の完全一致

組み合わせ

>> - シーケンス

```
# A の後に B が続く
rule(:word) { letter >> letter }
```

意味: 順序を持つ連結 (AND)

| - 選択

```
# A または B (順序が重要!)
rule(:token) do
  double_hyphen_word | # 先に試す
  hyphenated_word       # 後で試す
end
```

重要: PEG は最初にマッチした選択肢を採用

.repeat - 繰り返し

```
match['a-z'].repeat      # 0回以上
match['a-z'].repeat(1)   # 1回以上
```

AST 構築

.as(:symbol) - 命名

```
# Token に型を付ける
rule(:word) {
  letter.repeat(1).as(:word)
}
```

```
# 出力される AST
{ word: "hello" }
```

意味: AST で識別するための名前

root() - 開始規則

```
# Parser の入口を指定
rule(:sentence) {
  token >> space?
}
root(:sentence)
```

意味: どの規則から解析を始めるか指定

Regexp → Parslet への変換 (GSUB パターンから Parser 規則へ)

標点符号の処理

GSUB 方式

```

# 65+ パターンの一部
ROMAN_GSUB_PATTERNS = {
  '/,/ => '，'，',      # カンマ前後に空白
  '/./ => '。'，',      # ピリオド前後に空白
  '/!/' => '！'，',      # 感嘆符前後に空白
  '/?/' => '？'，',      # 疑問符前後に空白
  # ... 他 60+ パターン
}

# 適用
text = "suà-lòh,lâi-khuànn"
ROMAN_GSUB_PATTERNS.each do |pattern, replacement|
  text = text.gsub(pattern, replacement)
end
# => "suà-lòh , lâi-khuànn"

```

特徴: 記号を空白で囲む → 後で split

Parslet 方式

```
# 標点符号を Token として直接認識
rule(:punctuation) do
  str('...') | str('....') | str('.....') | # 複数文字優先
  match['.,.;()!?!/?/~^-_='] | # 単一文字
  match["\"'\u201C\u201D\u2018\u2019"] | # 引用符 ( いんようふ )
  match['\u3000-\u303F'] # CJK記号
end

# Token 規則
rule(:token) do
  hyphenated_word.as(:word) |
  punctuation.as(:punct)
end
```

入力: "suà-lòh, lâi-khuànn"

出力 (AST) :

```
[  
  { word: "suà-lóh" },  
  { punct: "," },  
  { word: "lái-khuànn" }  
]
```

特徴: Token として構造化 → split 不要

Regexp → Parslet への変換（連字符処理と音節数による漢字対応）

連字符の保持 (Page 17)

GSUB 方式

```
# Step 1: 記号正規化
text = "suà-lòh lái-khuànn"
# ハイフンは保持 (重要 !)

# Step 2: スペース分割
tokens = text.split(/\s/)
# => ["suà-lòh", "lái-khuànn"]

# Step 3: 音節数をカウント
syllables = "suà-lòh".split('-').size
# => 2

# Step 4: 漢字を音節数分取得
kanji_chars = ["繼", "落", "來", "看"]
combined = kanji_chars.shift(syllables).join
# => "繼落"
```

原理: ハイフン = 音節区切り

Parslet 方式

```
# 連字符付き単語を1つの Token として認識
rule(:hyphenated_word) do
  syllable >>
  (single_hyphen >> syllable).repeat
end

# "suà-lòh" → { word: "suà-lòh" }
```

音節数の計算:

```
# Phase 3: Transform
rule(word: simple(:w)) do
  syllables = w.to_s.split('-').size
  # => 2
end
```

漢字対応:

```
# 音節数 = 漢字文字数
"suà-lòh".split('-').size # => 2
"繼落".chars.size # => 2
# ✅ 一致!
```

原理: Parser が音節構造を保持 → 自動対応

Ruby Parser との比較(ひかく)

Ruby Parser (Prism)

```
# 入力  
"def foo(x); x + 1; end"
```

Phase 1: Lexical

```
[DEF] [IDENTIFIER] [LPAREN] [IDENTIFIER]  
[Rparen] [SEMICOLON] [IDENTIFIER] [PLUS]  
[INTEGER] [SEMICOLON] [END]
```

Phase 2: Syntax

```
DefNode(  
  name: :foo,  
  parameters: ParametersNode(...),  
  body: StatementsNode(...)  
)
```

Phase 3: Semantic

- Type checking
- Scope analysis
- Code generation

台羅 Parser (RomanParserPure)

```
# 入力  
"suà-lóh lâi-khuànn"
```

Phase 1: Lexical

```
[suà-lóh][lâi-khuànn]
```

Phase 2: Syntax

```
{  
  sentence: [  
    { word: "suà-lóh" },  
    { word: "lâi-khuànn" }  
  ]  
}
```

Phase 3: Semantic

- AST transformation
- Array generation

```
["suà-lóh", "lâi-khuànn"]
```

漢字処理は POJ Parser に依存

一方向の依存関係：複雑な構造を先に解析

POJ Parser (複雑)

```
# RomanParserPure - Parslet で実装
roman_array = [
  "suà-lóh",      # 2音節
  "lái-khuànn",   # 2音節
  "Sin-tik-tshí"  # 3音節
]

# 音節数の計算
"suà-lóh".split('-').size  # => 2
"Sin-tik-tshí".split('-').size # => 3
```

複雑な処理:

- ✓ ハイフンの意味解析（音節 vs 語間）
- ✓ 声調記号の認識（Unicode 結合文字）
- ✓ 二重ハイフン（--）の特殊処理
- ✓ 構文規則の定義と AST 構築

漢字処理（シンプル）

```
# POJ の音節数に従うだけ
kanji = "繼落來看新竹市"

# 1. "suà-lóh" = 2音節
#     → 漢字2文字: "繼落"
# 2. "lái-khuànn" = 2音節
#     → 漢字2文字: "來看"
# 3. "Sin-tik-tshí" = 3音節
#     → 漢字3文字: "新竹市"

kanji_array = ["繼落", "來看", "新竹市"]
```

シンプルな処理:

- ✓ 音節数 = 文字数の対応
- ✓ パターンマッチング（Edge Case 用）
- ✓ 独立した構文解析は不要

RomanParserPure の実装を試してみよう

GitHub で公開中：テストデータと検証スクリプト



<https://github.com/ryudoawaru/rwc2025-slide>

含まれているもの：

- 完全な RomanParserPure 実装
- 3000 件の実際のコーパスデータ

✨ テスト結果

```
$ ruby test_parser.rb
```

```
=====
```

```
Testing RomanParserPure with 3000 records
```

```
=====
```

```
[██████████] 100.0% (3000/3000)
```

```
=====
```

```
Final Results
```

```
=====
```

```
Total records: 3000  
Parse success: 3000 (100.0%)  
Parse errors: 0 (0.0%)
```

```
=====
```

```
🎉 PERFECT! 100% success rate achieved!
```

重要なポイント：

- ✅ 100% Parse 成功率 - 3000 件すべて正確に解析
- ✅ エラーなしで完全動作 - 実用性と理論の両立

プロジェクトの成果

Rubyで実現した台湾語教育システム

台灣語コーパスシステム

- **公式名:** 臺灣台語語料庫 應用檢索系統
- **公開URL:** <https://tggl.naer.edu.tw>
- **委託元:** 教育部（文部省相当）・国家教育研究院

The screenshot shows the homepage of the Taiwan Hakka Corpus Application Search System. At the top, there is a navigation bar with links for '臺灣台語語料庫應用檢索系統' (Homepage), '檢索系統' (Search System), '系統說明' (System Description), '語料申請' (Corpus Application), '相關連結' (Related Links), 'Open ID 登入' (Login), and user icons. Below the navigation bar, the title '臺灣台語語料庫 應用檢索系統' is displayed in large white text, with its Pinyin equivalent 'Tâi-uân Tâi-gí Gí-liāu-khò Ing-iōng Kiám-sik Hē-thóng' underneath. The background features a blurred image of bookshelves. Three main search functions are highlighted in boxes: '語料檢索' (Corpus Search) with a document icon, '教科書詞彙檢索' (Textbook Vocabulary Search) with a document and magnifying glass icon, and '語法點檢索' (Morphology Point Search) with a document and text icon. Below these boxes, a section titled '最新消息' (Latest News) contains a message about corpus applications being open: '語料申請已經正式開放囉!' (Corpus application has been officially opened!). The date '2024-06-19' is also shown. At the bottom, there is footer information for the National Academy for Educational Research, including address, phone number, fax, and links to privacy policy and feedback forms. A red logo for '5xRuby CO., LTD' is in the bottom right corner.

主要機能1：語料檢索

漢字・POJ・音声ファイルの統合検索システム

The screenshot shows the homepage of the "臺灣台語語料庫應用檢索系統". The top navigation bar includes links for "最新消息" (Latest News), "計畫介紹" (Project Introduction), "操作說明" (Operation Instructions), "語料申請" (Corpus Application), "相關連結" (Related Links), "Open ID 登入" (Open ID Login), and a search icon. Below the navigation is a banner with three tabs: "語料檢索" (Corpus Search), "教科書詞彙檢索" (Textbook Vocabulary Search), and "語法點檢索" (Grammatical Point Search). The main search area features input fields for "前字串" (Initial String) and "後字串" (Final String), both set to 10 characters. A search button labeled "檢索" (Search) is located to the right of the input fields. Below the search area is a section titled "語料檢索系統說明" (Corpus Search System Description) containing a detailed explanation of the system's features and purpose. At the bottom, there is a "收錄語料出處說明" (Source of Recorded Materials) section with two numbered points: 1. TATMOE: 收錄「教育部閩南語語音語料庫建置計畫」之成果，並有對應音檔可供播放及下載。 2. 教科書語料：收錄國小教科書之詞彙及例句，經真平企業有限公司授權使用。

主要機能 2: 教科書語彙

台湾の教科書で使用される台湾語語彙のデータベース

The screenshot shows the homepage of the "臺灣台語語料庫應用檢索系統". At the top, there are navigation links: 檢索系統, 系統說明, 語料申請, 相關連結, Open ID 登入. Below the header, there are three search tabs: 語料檢索, 教科書詞彙檢索 (highlighted in red), and 語法點檢索. A search form has two radio buttons: 台語漢字 (selected) and 臺羅. A search input field contains "輸入台語字詞查詢教科書詞彙" and a "檢索" button. Below the form, instructions are provided:

- 選擇查詢「台語漢字」或「臺羅」，輸入查詢字詞後，點選「檢索」。
- 查詢結果依完全符合、部分符合列出教科書語料中包含查詢字詞的詞語，並提供詞類、搭配詞、與《臺灣台語常用詞辭典》連結，可點選「看教科書語料」瀏覽完整句子。

查詢字詞之詞類1為教育部《臺灣台語常用詞辭典》收錄之詞性標注；詞類2為本研究參考「中研院平衡語料庫」46類簡化詞類標記，詳見「[中央研究院漢語料庫的內容與說明](#)」。

The footer of the website includes the logo of the National Academy for Educational Research (國家教育研究院). It also contains contact information for the Sanxia branch: 電話: (02) 7740-7890, 傳真: (02) 7740-1170, 地址: 237201 新北市三峽區三樹路2號. There are also links for 會員註冊及隱私聲明 and 意見回饋. The footer also features a QR code and the text "無障礙標章2.1".

國家教育研究院
National Academy for Educational Research

三峽總院區
電話: (02) 7740-7890
傳真: (02) 7740-1170
地址: 237201 新北市三峽區三樹路2號

個資法及隱私聲明
意見回饋
無障礙標章2.1

線上人數: 6 累積總人次: 1,633,557

34 / 37

© 2023 國家教育研究院. All Rights Reserved.

主要機能 3: 語法點檢索（文法ポイント検索）

台湾語の重要な文法パターンと例文の検索

The screenshot shows the search interface for grammatical points. At the top, there are tabs for '語料檢索' (Corpus Search), '教科書詞彙檢索' (Textbook Vocabulary Search), and '語法點檢索' (Grammatical Point Search). Below the tabs, there are two radio buttons: '台語漢字' (Hakka Chinese) and '臺羅' (Taleo). A search bar contains the placeholder '輸入台語字詞查語法點' (Input Hakka word to search for grammatical points) and a '檢索' (Search) button. Below the search bar, instructions provide guidance on how to use the feature. The main content area displays a table of grammatical points with columns for '編碼' (Code), '語法點' (Grammatical Point), '語法說明' (Grammatical Description), and '例句' (Example Sentences). Each row includes a '顯示筆數' (Number of results displayed) dropdown set to '20 筆' (20 items). The first four rows of the table are as follows:

編碼	語法點	語法說明	例句
001-01	是	即「是」。否定用「毋是」。	<ul style="list-style-type: none"> 這是真危險的代誌。 我攏是班裡的第一名。 看更多 ▾
002-01	是……的	強調「是……的」中間的成分。	<ul style="list-style-type: none"> 口味是一粒一的。 心肝頭實在是足疑的。 看更多 ▾
003-01	……的……	表示所有格、修飾語，「的」後面的成分可以省略。	<ul style="list-style-type: none"> 我的車幹來塗去， 伊的英語名是啥物？ 看更多 ▾
004-01	……的	置於名詞、形容詞、動詞之後，省略其後的名詞，指稱具備該特徵的名詞。	<ul style="list-style-type: none"> 阿爸畫的到底是啥物？ 閹豬的當做掠龍的。 看更多 ▾

結論

コンパイラ理論の普遍性

- プログラミング言語の Parser → 自然言語の処理
- Ruby の 3-phase 分析 → 台湾語の分詞アライメント

適切な道具を選び、原理を理解すれば、複雑な問題も解決できる

Conference から学び、新しい領域に挑戦する

- 既存の知識を新しい問題に応用
- エンジニアとしての成長の道



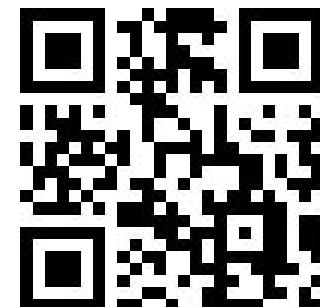
ご清聴ありがとうございました

スライド & コード



<https://github.com/ryudoawaru/rwc2025-slide>

5xRuby



<https://5ruby.com>

ブース出展中



ぜひお立ち寄りください！