Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

La science participative et l'ANR DiLSi

Pierre Magistry Ilaine Wang
ERTIM, Inalco, 2 rue de Lille, 75007 Paris, France
prenom.nom@inalco.fr

RÉSUMÉ

Cette communication propose un retour d'expérience sur les interactions entre le projet DiLSi et les communautés de locuteurs du teochew de la diaspora et du tâigí.

ABSTRACT

Citizen Science and the DiLSi ANR Project

This report offers an overview of the interactions between the DiLSi project and the speaker communities of diasporan Teochew and tâigí.

MOTS-CLÉS : science participative, langues peu dotées, oralité, variation, langues sinitiques.

KEYWORDS: citizen science, low-resource languages, orality, language variation, Sinitic languages.

ARTICLE : **Accepté à** ParCol 2025.

1 Introduction

L'ANR DiLSi s'intéresse à différents axes de variation des langues sinitiques et à leurs conséquences pour le TAL. Parmis ceux-ci, l'axe sur la variation diatopique s'inscrit dans les recherches sur langues peu dotées et minorées, en adoptant des pratiques de science participative. Les langues concernées sont le teochew et le tâigí, deux langues sinitiques proches, de la famille des langues dites « min du Sud ». Elles sont cependant dans des situations sociolinguistiques très différentes. Nous nous focalisons sur le teochew de la diaspora, tel que parlé en France (mais aussi en Amérique de Nord et en Asie du Sud-Est), tandis que le tâigí étudié est celui de Taïwan. Jusqu'à présent, les plus fortes interactions entrant dans le champ des sciences participatives ont été menées sur le teochew en liens étroits avec l'association des jeunes teochew de France (JTC), à Paris et sur Internet, notamment sur deux serveurs Discord : celui des JTC et le Discord *Gaginang* regroupant la diaspora internationale autour de celle des USA.

L'éloignement du terrain taïwanais contraste avec les interactions intenses sur le teochew. Dans le cadre de DiLSi, le tâigí est une langue proche, mieux outillée que le teochew et en voie de standardisation, ainsi qu'un cas d'étude pour le TAL ciblant les apprenants du tâigí en France.

Pour cette présentation, nous proposons d'insister principalement sur les interactions, coopérations ou contradictions entre un projet (ANR) académique et les projets communautaires et associatifs visant à étudier, transmettre, valoriser et outiller une langue d'héritage.

Actes de l'atelier Science Participative pour les Données et Corpus Linguistiques 2025 (ParCol), pages 3–5.

2 Agenda concerté

Notons d'abord qu'avant d'être déposé, le projet ANR a largement mûri après une longue expérience du terrain associatif de l'Open Data taïwanais, puis des ateliers *Contribuling*[1] organisés en partenariat entre ERTIM et Wikimedia, où la communauté teochew fut conviée et s'est montré très active.

Ces étapes préalables sont indispensables tant la rédaction d'un projet sur plusieurs années avec ses livrables et son diagramme de Gantt sont en contradiction directe avec une nécessaire négociation de l'agenda, des objectifs et des priorités entre la communauté concernée et le milieu académique responsable de la gestion d'un soutien logistique considérable (comparé aux moyens associatifs) mais limité dans le temps.

Une science pleinement participative doit laisser la possibilité d'infléchir les priorités du projet, sans pour autant renoncer aux livrables « promis à l'ANR ». Un exemple d'une inflexion acceptable et réussie fut de commencer les expériences en synthèse de la parole en ciblant la complétion des audio d'un dictionnaire sur mobile[2], alors que notre question de recherche initiale portait plus sur les effets de transfert du tâigí vers le teochew. À l'inverse, nos travaux sur un premier corpus arboré restent assez loin des préoccupations actuelles de la communauté.

3 Un continuum d'objectifs plus ou moins partagés

Ces réflexions sur les priorités nous conduisent à identifier différents types de chantiers, allant des questions scientifiques décorrélées des préoccupations de la communauté de locuteurs, aux tâches identifiées comme prioritaires par les associations de la communauté.

Ces tâches prioritaires ne nécessitent pas l'intervention du milieu académique, et sont abordées en toute autonomie par la communauté avec ses propres moyens. Le dictionnaire sur mobile sus-cité est un bon exemple de réalisation faite longtemps avant notre arrivée. D'un autre côté, les tâches et les préoccupations a priori *purement* scientifiques ne sont pas toutes sans conséquences sur les orientations de l'association, notamment pour tout ce qui se rapproche de l'éducation populaire. Par exemple, nos propres questionnements poussent en effet vers des discussions sur le statut de la langue et sa revalorisation par la pratique de l'écrit. Ont été mis en place des ateliers de transcription en *peng'im* et de prise en main d'un clavier virtuel que nous avons créé. Nos travaux encouragent également un traitement éclairé de la variation intradialectale que l'on cherche à documenter.

4 Réalisations

Corpus oral d'enregistrements en teochew (monologues, dialogues)
Corpus arboré trilingue tâigí-teochew-mandarin
Synthèse vocale pour le teochew pour compléter un dictionnaire sur mobile
Méthodes de saisie pour écrire le teochew en sinogrammes et romanisation sur mobile
Reconnaissance vocale et travail sur la diversité des accents

Références

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

1. Trois éditions ont été organisées entre 2021 et 2023 (voir par exemple la page de l'édition 2023 : https://meta.wikimedia.org/wiki/ContribuLing_2023).
2. *WhatTCSay3*, dictionnaire mobile anglais-teochew et français-teochew disponible sur Apple Store et Play Store.