

Johanna Cordova

Le quechua dans les outils numériques, un défi pour le TAL ?
Développement de ressources linguistiques et numériques pour le
quechua ancashino

Directeur de thèse : César Itier, Damien Nouvel

Date de soutenance : 18 décembre 2024

Résumé

Les langues quechuas constituent l'une des familles linguistiques amérindiennes comptant le plus grand nombre de locuteurs natifs. Au Pérou, selon le recensement de 2017, 13,9% de la population a le quechua pour première langue et environ 20% le parle. Pourtant, elle est presque totalement absente des usages numériques. En traitement automatique des langues (TAL), c'est une langue peu dotée, avec une forte disparité de ressources selon la variété de quechua considérée. L'objectif de cette thèse est de développer un ensemble d'outils fondamentaux pour le traitement automatique d'une variété du quechua central, le quechua ancashino, parlé par environ 400 000 personnes, et en danger d'extinction d'après la classification de l'UNESCO. Ce processus comporte trois étapes : la numérisation des ressources disponibles dans cette variété (dictionnaires, corpus écrits), l'implémentation d'un analyseur morphologique, et l'élaboration d'un corpus arboré pour l'analyse en morphe-syntaxe. Les ressources développées seront valorisées à travers des applications telles qu'un moteur de recherche permettant d'interroger l'ensemble des dictionnaires. Dans un contexte global de valorisation des langues originaires et alors que d'ambitieuses politiques liées aux droits linguistiques sont en cours de déploiement dans les pays de l'aire andine, la présence du quechua dans les technologies constitue un important levier pour renforcer sa pratique et faciliter son enseignement.