# Tractament automatic de l'occitan : l'escomesa de la variacion

Clamença Poujade[1]

[1]CLLE-ERSS, Universitat Jean-Jaurès - Tolosa 2

L'occitan es una lenga considerada coma pauc dotada dins lo domeni del tractament automatic de las lengas (TAL) naturalas. Aquò vòl dire que i a pauques utisses e paucas ressorgas disponiblas per ameliorar son tractament automatic. Dempuèi un desenat d'annadas, de grops de trabalh se forman, de còps en s'associant amb d'autras lengas pauc dotadas, per construsir d'utisses e de ressorgas pel TAL occitan (Bernhard et al. 2021 ; Bras et Vergez-Couret 2016 ; Miletic et al. 2020 ; Vergez-Couret et Urieli 2015). D'unes òbran sus l'oral (reconeishença vocala, sintèsi vocala, etc.), d'autres sus l'escrit (traduccion automatica, reconeishença de mots, etc.). Per l'occitan, la granda part de las recèrcas se fan amb la grafia classica. Atal, lo TAL de l'occitan es pas encara robuste fàcias a las variacions graficas, ni mai a d'unas variacions lingüisticas.

Dins aquela presentacion parlarai de ma recèrca de tèsi, aquela recèrca se concentra sus l'amelioracion e la creacion d'utisses de TAL per l'anotacion morfosintaxica automatica de tèxtes occitans per que pòscan anotar tèxtes de tota grafia e de tot dialècte. Per aquel trabalh avèm causit de trabalhar suls parlars occitans d'Arièja, per çò que aquel territòri conten una granda variacion lingüistica, amb las nombrosas isoglòssas que i passan. Los tèxtes qu'avèm recampats presentan una importanta variacion grafica (grafias classicas, mistralencas e non classicas). Mas l'ambicion de la tèsi es d'aver una metodologia reproductibla per tot dialècte, tota grafia e tota lenga pauc dotada, pas solament pels parlars d'Arièja o per l'occitan.

Presentarai totes los còrpus constituits per aquel trabalh, los objectius, la metodologia mesa en plaça e los resultats obtenguts.

Tot aquel trabalh sus l'anotacion de tèxtes a per tòca de permetre una anotacion automatica fisabla e atal ajudar lo desvolopament d'autras aisinas de TAL e tanben lo trabalh de recèrca sus l'occitan. Permetrà, egalament, la creacion o l'amelioracion d'aisinas mai viradas cap al grand public per exemple, l'amelioracion de traductors automatics, de sintèsi vocala, etc.

# Referéncias


Bernhard, D. et al. (2021). « Collecting and annotating corpora for three under-resourced languages of France : Methodological issues ». In : *Language Documentation & Conservation* 15, p. 316-357. URL : https://hal.archives-ouvertes.fr/hal-03273196.

Bras, Miriam et Marianne Vergez-Couret (2016). « BaTelÒc : A text base for the Occitan language ». In : *Language Documentation & Conservation : Language Documentation and Conservation in Europe. Special Publication No. 9*. Sous la dir. de Vera Ferreira et Peter Bouda, p. 133-149.

Miletic, A. et al. (2020). « Building a Universal Dependencies Treebank for Occitan ». In : *12th Language Resources and Evaluation Conference*, p. 2932-2939. URL : https://hal.archives-ouvertes.fr/hal-02892715.

Vergez-Couret, Marianne et A. Urieli (2015). « Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan ». In : *TALARE 2015*. URL : https://hal.archives-ouvertes.fr/hal-01214566.