PROPOSITION DE STAGE

Année 2019

Laboratoire L3i

Sujet de stage :

Localisation, reconnaissance et extraction des données hétérogènes dans des documents cartographiques

Résumé du travail proposé :

L'objectif du stage est de savoir localiser, reconnaitre et extraire les différents types d'informations qui se trouvent dans l'Atlas linguistique de France. Ce travail sera un travail en lien avec une thèse effectuée au laboratoire L3i afin d'approfondir le travail commencé. Si la partie segmentation est terminée, le stagiaire pourra s'essayer à l'implémentation des méthodes de reconnaissance de l'état de l'art.

Mots clés :

Traitement de l'image, Morphologie mathématique, Composantes connexes, Analyse de carte, Séparation texte/graphique, Atlas linguistique

Informations complémentaires :

Encadrant(s) : **Jordan DRAPEAU**

Equipe :

- ☒ **Images et Contenus**
- ☐ **Dynamique des systèmes et adaptativité**
- ☐ **Modèle et Connaissance**

Domaine d'application stratégique :

- ☐ **E-éducation**
- ☐ **Environnement et développement durable**
- ☐ **E-culture**
- ☒ **Valorisation de contenus numériques**

Cadre de coopération : **ANR ECLATS**

Date de début du stage : **01/02/2019**

Durée du stage : 6 mois

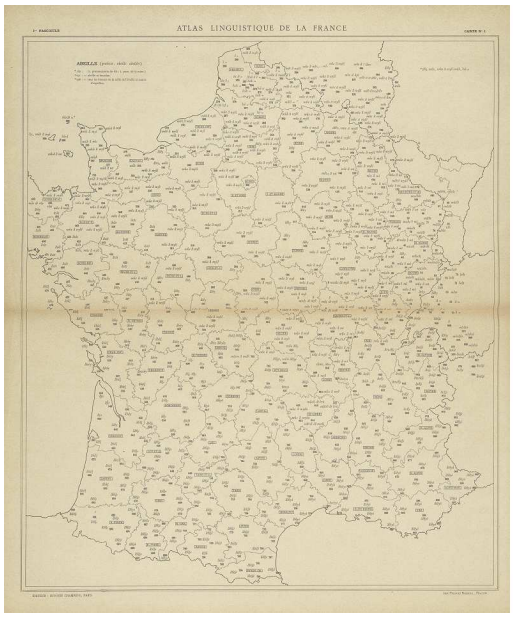
Financement : ANR ECLATS

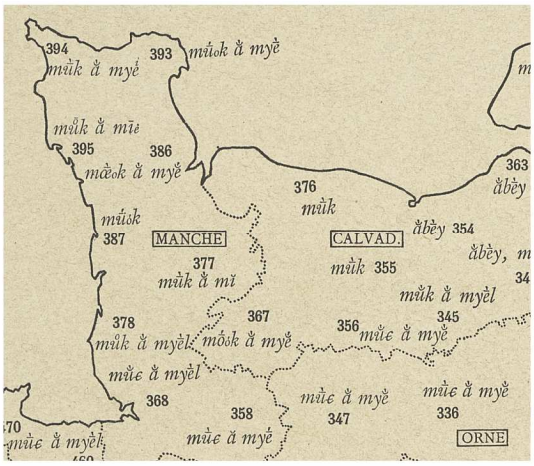
Contexte de l'étude:

Ce travail rentre dans le cadre du projet ANR ECLATS. Ce projet concerne la valorisation et l'analyse des documents cartographiques anciens, un patrimoine historique et culturel reconnu comme source d'information particulièrement riche mais difficilement exploitable. Celui-ci s'intéresse plus particulièrement à l'Atlas Linguistiques de France (ALF), élaborés entre 1902 et 1910 et qui fournit les données de premier ordre en dialectologie. L'objectif est d'apporter un outillage logiciel et méthodologique facilitant l'extraction, l'analyse, la visualisation et la diffusion

des données contenues dans les atlas linguistiques anciens afin de permettre des recherches novatrices en dialectologie.

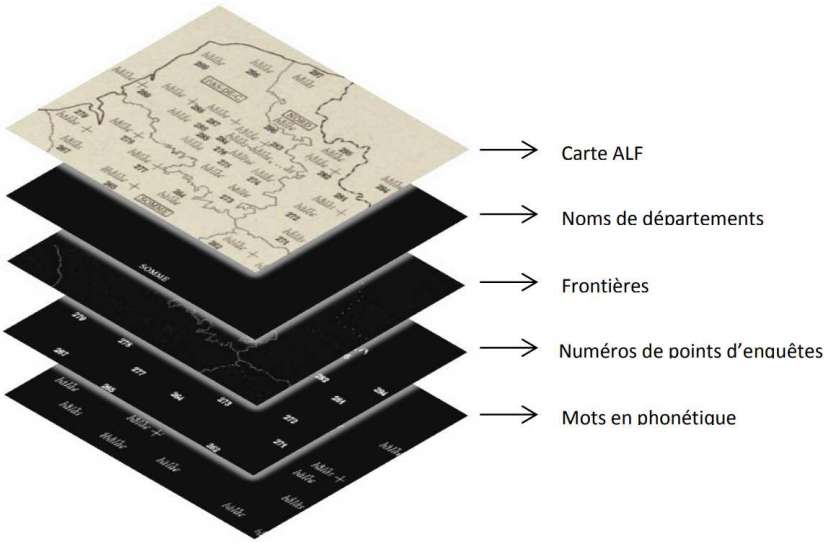
L'Atlas Linguistique de France (ALF) est une collection de 1920 cartes qui se regroupent dans une trentaine de fascicules, réunit en 12 volumes. Chaque carte de cet atlas fait référence à un seul mot sur laquelle figure toutes les formes dialectales transcrites phonétiquement, réparti sur la France et quelques villes des pays frontaliers. Quatre types d'informations figurent sur ces cartes : noms de département, numéros de point d'enquête, frontières, prononciations en phonétique du mot écrits en alphabet Rousselot-Gilliéron.





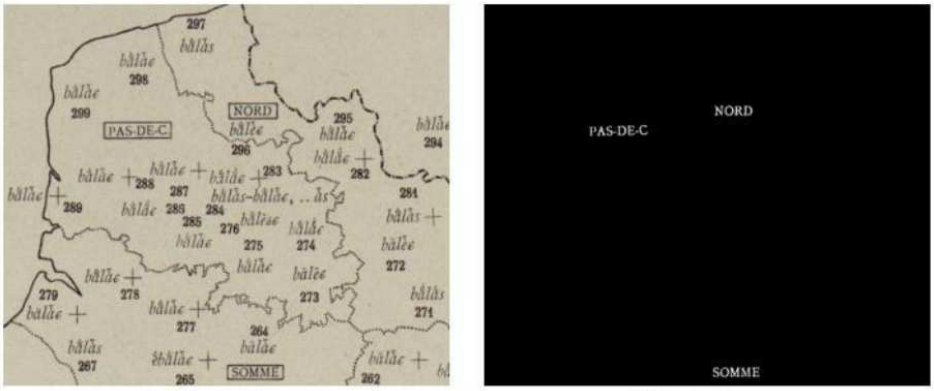
Description du sujet :

L'objectif du stage est de savoir localiser, reconnaitre et extraire en couche les différents types d'informations qui se trouvent dans l'Atlas linguistique de France.



Les pistes qui ont déjà été exploitées sont : les différents traitements basiques de l'image à l'aide de la morphologie mathématique, le filtrage d'arbres de composantes connexes ou encore la clusterisation des composantes connexes de ces cartes.

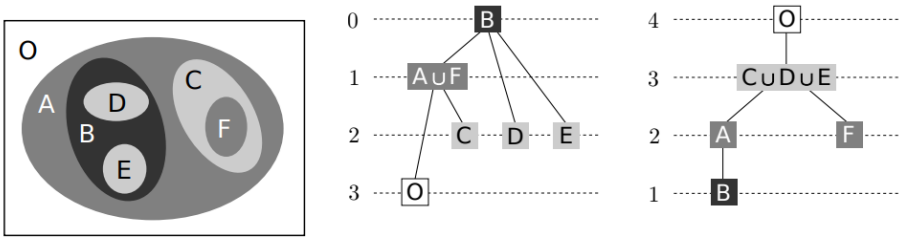
En ce qui concerne le traitement de l'image, une application à base d'outils très simples et courants (otsu, ouverture, fermeture, etc.) a été mise en place, mais travailler sur des images très grandes demandait des temps de calculs extrêmement long.



(a) a part of map (b) layer of names

(c) layer of numbers (d) layer of phonetics

Le plus gros du travail a été centré sur les arbres de composantes connexes. Il existe plusieurs types d'arbres de composantes connexes, mais deux particulièrement vont nous intéresser : le max-tree et le min-tree. Ce sont deux arbres qui vont chercher à classer toutes les composantes connexes de l'image donnée en entrée, suivant leur niveau de gris. Pour le max-tree, le nœud racine (root) est le nœud le plus sombre, et plus on va descendre dans les fils de l'arbre, plus les composantes seront claires, et inversement pour le min-tree, la racine est la composante la plus claire, et plus on descend dans les fils de cet arbre, plus les fils sont foncés. Une fois notre arbre fait, il nous suffit de filtrer les feuilles de l'arbre (composantes connexes) qui ne correspondent pas à ce qu'on cherche à obtenir à la fin de ce filtrage.



(a) image (b) max-tree (c) min-tree

Il faut que le programme soit robuste aux différents défauts présents sur les cartes. Les cartes de l'ALF sont anciennes et elles ont parfois eu du mal à être scannées parfaitement en raison de leur mode de reliure ou de leur état de dégradation.

Les cartes de l'ALF peuvent parfois atteindre une résolution assez importante, de l'ordre de 9704px par 11824px, soit environ 115 mégapixels. Il faut donc que l'application soit en mesure de supporter toutes les tailles d'images, surtout les plus grandes et cela dans des délais de traitement raisonnable.

Prérequis et contraintes particulières :

A la fin du stage, le travail effectué durant celui-ci devra permettre, par la séparation en couche d'informations, de mettre en place un système de reconnaissance des caractères

phonétiques en n'utilisant que la couche de mots phonétiques. Les autres couches serviront à replacer correctement les informations sur une carte refaite pour une interface web.

Références bibliographiques :

Jordan Drapeau, Thierry Géraud, Mickaël Coustaty, Joseph Chazalon, Jean-Christophe Burie, et al.. Extraction of ancient maps content by using trees of connected components. IAPR International Workshop on Graphics Recognition at ICDAR 2017, Nov 2017, Kyoto, Japan.

Contacts – liens :

Email :
jordan.drapeau@univ-lr.fr
jcburie@univ-lr.fr
mickael.coustaty@univ-lr.fr

ALF :
http://lig-tdcge.imag.fr/cartodialect3/