PROPOSITION DE STAGE

Année 2016

Laboratoire L3i

Sujet de stage :

Analyse de cartes linguistiques : Segmentation et reconnaissance d'éléments linguistiques

Résumé du travail proposé :

Le stage sera réalisé dans le cadre du projet ANR ECLATS. Ce projet a pour objectif de contribuer à la valorisation et l'analyse des documents cartographiques anciens. Il s'intéresse plus particulièrement à l'Atlas Linguistiques de la France (ALF).

Un travail préalable a permis de séparer les éléments graphiques et textuels présents dans ces cartes. L'objectif du stage sera dans un premier temps de segmenter les zones textuelles afin d'extraire les caractères phonétiques et les caractères « imprimés ». Dans un second temps, des algorithmes de reconnaissance de caractères seront testés pour évaluer leur capacité à reconnaitre les différents éléments présents sur les cartes.

Le travail sera réalisé en collaboration avec le laboratoire LIRIS de Lyon ainsi que le LIG et le GIPSA Lab de Grenoble.

Mots clés :

Traitements d'images, Segmentation, reconnaissance de texte, symboles et éléments phonétiques

Informations complémentaires :

Encadrant(s) : Mickael Coustaty ; Jean-Christophe Burie, Jean-Marc Ogier

Thématiques :

- ☐ **Ingénierie des connaissances**
- ■ **Analyse et gestions de contenus**
- ☐ **Interactivité et dynamique des systèmes**

Domaine d'application :

- ■ **Pertinence – contenu – interactions**
- ☐ **Environnement**

Cadre de coopération : Projet ANR ECLATS (en collaboration avec les laboratoires LIG – Grenoble, LIRIS – Lyon)

Date de début du stage : 4 janvier 2014

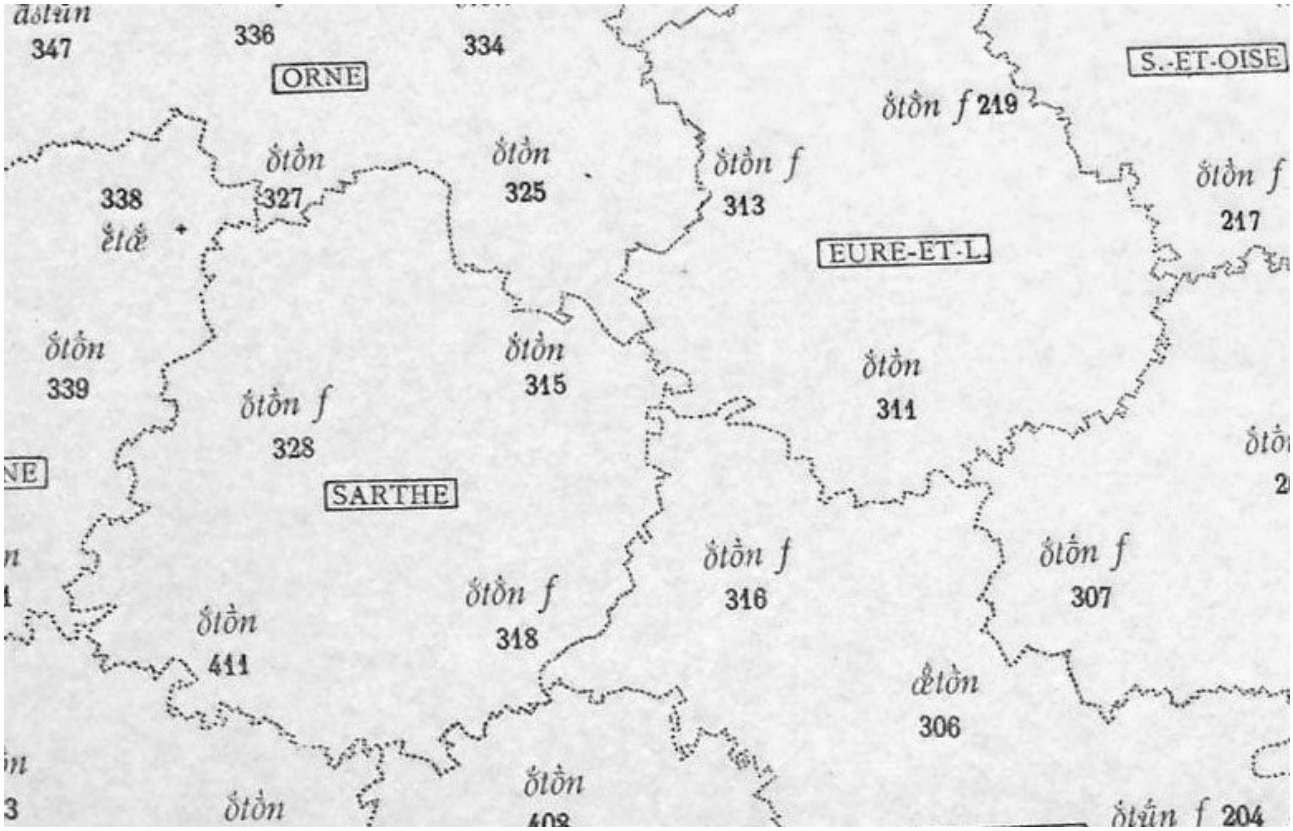
Durée du stage : 6 mois

Financement : ANR ECLATS

Contexte de l'étude:

La dialectologie s'intéresse à l'étude des traits linguistiques caractéristiques des langues à tradition orales comme les parlers locaux, appelés patois ou, encore, dialectes. Ces traits linguistiques peuvent être de nature très différente - phonétique, morphosyntaxique, lexicale, sémantique ou prosodique - et évoluent dans un espace géographique donné, dans le temps et au contact de la

société. Pour étudier les parlers locaux, la dialectologie s'est spécialisée dans la constitution de corpus de données descriptives, collectées via une méthodologie d'enquête qui repose sur des questionnaires, sur le choix de réseaux de points linguistiques et d'informateurs. Le traitement et l'analyse des données de terrain se fait au moyen de supports cartographiques, sur lesquels sont portées sous forme de points les localités enquêtées et les formes linguistiques collectées en transcription phonétique. À chaque concept exprimé par une entrée lexicale sous la forme d'un titre de carte en français est associée une et une seule carte sur laquelle figurent toutes les formes dialectales transcrite phonétiquement désignant le concept en question. L'Atlas Linguistique de France fut le premier du genre : après une campagne de terrain menée entre 1897 et 1901 dans 639 localités, J. Gilliéron et E. Edmont publient les résultats de 1902 à 1910. L'ALF concerne le domaine des dialectes gallo-romans de France, ainsi que d'une partie de la Belgique, de la Suisse, de l'Italie, et déborde également sur le domaine catalan de France et, dans un volume spécifique, sur la Corse. La publication de l'œuvre, en format papier, comporte 35 fascicules, réunissant en 12 volumes, 1920 cartes géolinguistiques qui présentent une photo instantanée de la situation dialectale de la France à la fin du XIXe siècle.



Extrait d'une carte de l'ALF : prononciation du mot « automne »

L'objectif du projet ECLATS est de développer des outils d'analyse automatique de ces cartes géolinguistiques afin de nourrir un système d'information géographie. Les données pourront ainsi être étudiées de façon plus efficace par les linguistes et dialectologue pour comprendre l'évolution de la langue.

Description du sujet :

L'objectif du stage sera de développer des outils de segmentation permettant d'extraire toutes les données textuelles : chiffre (points d'enquête), les départements, les symboles phonétiques en prenant soin de conserver tous les diacritiques (éléments situés généralement au-dessus des symboles). Ces petits éléments sont très importants car ils modifient la façon dont les mots sont prononcés. Les images étant parfois « bruitées » ou de qualité moyenne, les méthodes devront être robustes à ces perturbations.
Après avoir extraits les données textuelles, il faudra les classer selon les 3 catégories évoquées précédemment. Sur chacune de ces catégories, des techniques classiques de reconnaissances de

caractères (OCR) seront appliquées. L'idée est d'évaluer les performances de ces méthodes sur ce type de caractères. Les OCR tels que Tesseract, Abby Fine Reader, … pourront être utilisés.
En fonction de l'avancement du stage, les approches de « word spotting » qui permettent de localiser des suites de caractères identiques pourront également être testées sur les cartes de l'ALF.

Ce travail sera réalisé dans le cadre d'un projet interdisciplinaire. Des échanges avec des linguistes et des dialectologues pourront avoir lieu pour bien comprendre les caractéristiques de l'alphabet phonétique utilisé. En fonction de l'état d'avancement du projet, le stagiaire sera également amener à participer aux réunions pour faire part de son travail à l'ensemble des partenaires du projet.

Prérequis et contraintes particulières :

Le candidat doit :

- être actuellement en master 2 d'informatique ou justifier de compétences équivalentes
- de préférence, avoir suivi quelques cours de traitement et d'analyse d'images
- avoir un certain goût pour la recherche (étude bibliographique à prévoir)
- avoir un bon niveau de programmation
- avoir un bon niveau d'anglais lu et écrit

Le candidat intéressé enverra une lettre de motivation, ainsi qu'un CV détaillé, aux deux encadrants du stage, avant le 10 Novembre 2015. Des auditions seront organisées la seconde quinzaine du mois de novembre.

Références bibliographiques :

- Dang Quoc B., Muzzamil L. M., Coustaty M., Nayef N., De Tran C., Ogier J.-M. (2014), A multi-layer approach for camera-based complex map image retrieval and spotting system, IEEE IPTA 2014
- Guyomard J., Thome N., Cord M., Artières T.(2012), Contextual Detection of Drawn Symbols in Old Maps, in International Conference on Image Processing, ICI 2012,Orlando, 837-840
- Luo H, Agam G., Dinstein I. (1995), Directional Mathematical Morphology Approach for Line Thinning and extraction of Character Strings from Maps and Lines Drawings, in Proc. ICDAR'95, Montreal, Canada :257-260.

Contacts :

Email : mcoustat@univ-lr.fr, jcburie@univ-lr.fr