Les VOCAUX

Sciences participatives et nouvelles données pour la recherche en linguistique

Vendredi 30 Septembre 2022
Laboratoire ATILF, Nancy, salle Imbs

14h30 - 14h45	**Accueil**
14h45 - 15h15	**Un entrepôt de données langagières : ORTOLANG** Cyril Pestel (ATILF, CNRS-Université de Lorraine)
15h15 - 16h00	**L'appli Français de nos Régions : premiers résultats** Mathieu Avanzi (Université de Neuchâtel)
16h00 - 16h45	**Twitter et Reddit pour la recherche sociolinguistique quantitative : atouts, défis et solutions** Marie Flesch (ATILF, CNRS-Université de Lorraine)
16h45 - 17h30	**Le projet Les Vocaux : bilan d'étape** Julie Glikman (Université de Strasbourg, LiLPa & ATILF), Nicolas Mazziotta (Université de Liège, Traverse), Camille Fauth (Université de Strasbourg, LiLPa), Christophe Benzitoun (Université de Lorraine, ATILF)
17h30 - 17h45	**Clôture et bilan de la journée**

Accès : https://www.atilf.fr/contact/#acces

La salle Imbs se trouve au 2e étage.

Résumés ci-dessous.

Séminaire organisé par Julie Glikman et Christophe Benzitoun dans le cadre du projet **Oralité et Diachronie : Une voie d'accès aux formes émergentes,** financé par l'Université de Lorraine, le laboratoire ATILF (CNRS-U. Lorraine) et le laboratoire LiLPa (Université de Strasbourg).

Un entrepôt de données langagières : ORTOLANG

Cyril Pestel (ATILF, CNRS-Université de Lorraine)

Dans cette communication, nous présenterons la plateforme ORTOLANG. ORTOLANG est un service spécialisé pour la langue, complémentaire de l'offre générale proposée par la TGIR Huma-Num (Très Grande Infrastructure de Recherche). L'EquipEX est une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue française et son traitement clairement disponibles et documentés.

L'appli Français de nos Régions : premiers résultats

Mathieu Avanzi (Université de Neuchâtel)

La méthode du crowdsourcing (on dit aussi production participative, externalisation ouverte ou myriadisation) appliquée aux domaines des sciences (on parle alors de sciences participatives, sciences citoyennes ou sciences collaboratives) consiste à faire appel à la foule en vue de réaliser certaines tâches. Si la méthode ne date pas d'aujourd'hui, elle a tout récemment pris une certaine importance avec la démocratisation des smartphones et autres supports mobiles connectés. Dans le domaine de la géographie linguistique, ces développements technologiques ont offert des perspectives intéressantes en ce qui concerne le recueil de données à grande échelle. Au cours des cinq dernières années, des applications conçues pour être exclusivement téléchargées sur des supports connectés, tels que les smartphones et les tablettes, ont été conçues pour l'allemand et ses dialectes, l'anglais américain et l'anglais britannique, ainsi que le frison et le luxembourgeois. Ces applications comportent différents volets ludiques, destinés à appâter les utilisateurs (service de localisation de l'utilisateur, informations sur la voix ou son débit, etc.), mais leur intérêt réside surtout dans le fait qu'elles rendent possible l'enregistrement de données audio (parole lue, élicitée à partir d'images ou de façon moins contrainte, etc.), de façon simultanée, aux quatre coins des pays où elles sont diffusées, et ce avec une qualité acoustique qui permet des études instrumentales très fiables.

Dans cette communication, je présenterai la méthodologie que nous avons suivie pour mettre au point la première application mobile destinée à documenter le français que l'on parle d'un bout à l'autre de la francophonie. Je parlerai des enquêtes préliminaires que nous avons conduites dans le cadre du programme de recherche Français de nos Régions, puis des premiers résultats que nous livrent l'application, près de 6 mois après sa mise en ligne.

Twitter et Reddit pour la recherche sociolinguistique quantitative : atouts, défis et solutions

Marie Flesch (ATILF, CNRS-Université de Lorraine)

Cette intervention présentera des réflexions méthodologiques qui sont le fruit de la construction de grands corpus composés de tweets et de commentaires publiés sur Reddit. Après avoir mis en avant l'intérêt de ces données pour la recherche sociolinguistique, elle explorera les spécificités de Twitter, réseau social qui fournit souvent de riches informations sociodémographiques, et celles de Reddit, site communautaire sur lequel le pseudonymat est la règle. Elle s'attardera sur les défis posés par le recueil automatique de données sociodémographiques ; pour pallier les erreurs d'annotation et les problèmes éthiques, elle proposera une méthode qui allie techniques automatiques et exploration manuelle des données.

Le projet Les Vocaux : bilan d'étape

Julie Glikman (Université de Strasbourg, LiLPa & ATILF), Nicolas Mazziotta (Université de Liège, Traverse), Camille Fauth (Université de Strasbourg, LiLPa), Christophe Benzitoun (Université de Lorraine, ATILF)

Ces dernières années ont vu le développement d'un nouveau mode de communication sous la forme de messages vocaux enregistrés, appelés « sms vocaux », « notes vocales » ou encore « vocaux », terme que nous retiendrons ici, diffusés sous forme de messages, à destinataires uniques ou multiples, connus ou inconnus, via des plateformes de type Snapchat, Instagram, ou des logiciels de messagerie type WhatsApp ou Messenger. Véritables « sms vocaux/vidéos », ils se substituent purement et simplement à l'envoi de sms, sans pour autant constituer un retour à la conversation téléphonique. Ainsi, ces productions seraient le pendant, en code oral, du sms écrit, dont l'étude s'est développé grâce au projet *sms4science* (http://www.sms4science.org/, voir Panckhurst et al. 2013 pour le corpus *88milSMS* recueilli à Montpellier) ou grâce au projet *What's up, Switzerland* (Stark (2016-2018) www.whatsup-switzerland.ch).

Le projet Les Vocaux a pour objectif d'étudier ce nouveau mode, à travers la constitution d'un corpus inédit de « vocaux ». L'objectif de notre communication sera de montrer l'intérêt pour la recherche que représentent de telles données, ainsi que de faire un bilan d'étape sur la constitution du corpus, du recueil au format final de distribution.