"Les données représentent le coeur de mon travail" Thomas Gaillat

Travaillant sur un gros corpus de texte et d'enregistrement d'entretiens, Thomas Gaillat présente comment il collecte et mets à disposition ses données dans l'objectif que d'autres chercheurs puissent les manipuler et effectuer des tâches de modélisation, et dans le respect de la vie privée.

Transcription textuelle de la vidéo

Interview de Thomas Gaillat, enseignant-chercheur en anglais

"Je suis Thomas Gaillat, enseignant-chercheur en linguistique et didactique des langues, rattaché à l'équipe LIDILE de l'Université de Rennes 2. Par ailleurs, je suis enseignant d'anglais, de spécialité, au centre de langue de l'Université de Rennes 2.

Mes questions de recherche tournent autour de la problématique de l'apprentissage des langues, à savoir comment est-ce que les gens apprennent une langue, par quelles étapes ils passent, elles passent pour acquérir différents schémas, différentes structures linguistiques. Pour faire ce travail nous reposons sur la collecte de données. Il nous faut des données textuelles, il faut que les apprenants, qu'on appelle les étudiants aussi par ailleurs, nous collectons des textes, des enregistrements de ces sujets et ensuite il faut pouvoir les traiter automatiquement parce que c'est aussi un des enjeux, c'est de passer à l'aspect quantitatif des choses et donc de audio pour ensuite permettre d'appliquer des méthodes de modélisation statistiques de manière à extraire un certain nombre de variables qui sont pertinentes dans l'explication du pourquoi les gens utilisent telle ou telle structure.

Et donc ces données-là, et c'est dans ce cadre-là que la science ouverte intervient, sont hyper importantes puisque pour avoir des bons systèmes de modélisation, des bons modèles, il nous faut beaucoup de données et des données de qualité. Alors dans ce cadre-là, à l'équipe Lidile depuis maintenant plus près de deux décennies, il y a la collecte d'un corpus de français langue étrangère et d'anglais langue étrangère. Donc l'équipe depuis maintenant 2008 me semble collecte ce type de données, l'équipe a le plaisir de pouvoir numériser ces données, elle l'était déjà, mais surtout de les mettre à disposition sur une base de données humanum, Nakala, et la communauté scientifique peut maintenant se connecter sur cette base, soit par interface humaine, soit par interface informatique et des programmes peuvent se connecter directement sur les différents enregistrements, les différentes retranscriptions pour en extraire et les passer dans de nouvelles chaînes de traitement. Et donc quand on constitue un corpus, et c'est mon dernier point, les chercheurs pour arriver vers une science ouverte, c'est-à-dire déployer des corpus, les rendre en accès libre, il y a des locuteurs que nous enregistrons. Et c'est d'autant plus important, par exemple, si on recueille des données sur des apprenants qui sont à Taïwan, et si dans des entretiens semi-guidés ils sont amenés à donner leurs opinions politiques sur, disons, leurs voisins, on peut comprendre que parfois ça pourrait, ou dans le futur, ça pourrait poser des problèmes.

Donc ils font accompagner les enseignants dans comment je rend les données anonymes, comment je les rends compatibles avec la GRDP. Et pour ça, justement, j'ai collaboré avec le service SOCLE parce que l'ensemble de ces données et la façon dont on les rend disponible et la façon dont on informe les apprenants de la manière dont elles sont disponibles et ce qu'elles peuvent en faire par la suite nécessite une formation pour les enseignants-chercheurs qui doivent se préoccuper des plans de gestion des données, comment ces données sont décrites et mises à disposition."

25 octobre 2021