Traitement automatique des langues régionales de France : retour d'expérience sur les dialectes alsaciens

Delphine Bernhard, Université de Strasbourg, LiLPa, UR 1339

Les langues dites « peu dotées » sont souvent négligées dans les travaux de recherche en traitement automatique des langues (TAL), au profit des langues très dotées. Ces dernières disposent de nombreuses ressources linguistiques numériques, et en particulier de données annotées. Par exemple, une recherche avec le terme générique « corpus » sur le portail européen META-SHARE montre que plus du quart des ressources listées sont en anglais[1].

On assiste toutefois à une timide évolution dans la manière de traiter et considérer la diversité linguistique en TAL, grâce à des ateliers dédiés comme CCURL (*Collaboration and Computing for Under-Resourced Languages*), des groupes de travail (SIGUL : *ISCA Special Interest Group on Under-resourced Languages*, LITHME : *Language in the Human-Machine Era*) ou des projets, notamment européens (DLDP : *Digital Language Diversity Project*, ELE : *European Language Equality*).

Dans cette communication, nous nous intéresserons plus particulièrement au cas des dialectes alsaciens et aux progrès réalisés ces dernières années, grâce à divers projets. Nous nous appuierons notamment sur les résultats du projet RESTAURE [2] (Ressources informatisées et Traitement Automatique pour les langues régionales), financé par l'ANR (2015-2018) et dédié à trois langues régionales de France : l'alsacien, l'occitan et le picard. Nous insisterons plus particulièrement sur les défis posés : manque de données numériques, variations dialectales et graphiques, communauté de recherche réduite, difficultés à valoriser le travail de collecte et d'annotation des données. Nous montrerons également quelles solutions méthodologiques peuvent être privilégiées pour faciliter la création de ressources et renforcer la visibilité de ces langues : coopération entre équipes de recherche intéressées par diverses langues régionales, collaboration de chercheurs et chercheuses appartenant à diverses disciplines, utilisation de standards, réutilisation d'outils existant, adoption des principes FAIR[3] pour la diffusion des ressources. Il s'agira ainsi de montrer comment le travail sur des langues régionales peu ou très faiblement dotées peut, au-delà des réalisations et productions concrètes, enrichir et alimenter la réflexion sur les pratiques de recherche en TAL et linguistique outillée.

[1] http://metashare.ilsp.gr:8080/repository/search/?q=corpus Recherche effectuée le 17/05/2021.
[2] https://restaure.unistra.fr
[3] *Findability, Accessibility, Interoperabily and Reuse*. Voir Wilkinson et al. (2016) « The FAIR Guiding Principles for scientific data management and stewarship », Scientific Data, 3, https://www.nature.com/articles/sdata201618