« Constitution de corpus et outils » : notation/annotation ; variation/norme.

La diversité des sources de données linguistiques soulève quantité de problèmes liés à la nécessité de passer de « données collectées » (ouvrages numérisés, enregistrements de données orales ou multimodales) à des « données de la recherche ». Le passage de l'un à l'autre suppose des opérations de transcription, de *notation*. On distingue classiquement le « corpus brut » ainsi produit du corpus enrichi d'*annotations*, informations associées au corpus brut pour en assurer l'exploitabilité. L'*annotation* est une composante heuristique de la recherche, dans la mesure où elle procède à des catégorisations qui permettront de filtrer les données lors de la constitution de corpus d'études, puis de se livrer à l'étude des observables sélectionnés.

Cette distinction entre *notation* et *annotation* demande aujourd'hui à être problématisée, les *notations* sont le résultat de choix faits en respectant des protocoles, et s'apparentent donc à une première phase d'*annotation*, les *annotations* quant à elles, sont analysées et exploitées comme des *données* sur lesquelles s'appuie la recherche. Notations et annotations répondent à un processus de **normalisation** permettant d'organiser l'infinie variation des observables linguistiques : le dialogue entre *notations* et *annotations* rejoint donc celui entre *norme* et *variation*. Se pose aussi la question de l'annotation manuelle de corpus (et des outils y contribuant), la catégorisation expérimentale sur des corpus de diverses tailles étant au cœur de l'analyse modélisante des observables linguistiques.