Mate-shs – atelier données textuelles des JA 2019

Cet atelier a révélé une grande diversité de profils parmi les participants, et une forte disparité en raison de la variété des « traitements » pratiqués et surtout de l'origine des participants, c'est-à-dire la manière dont ils/elles ont commencé à aborder les données textuelles. Si certain(e)s ont abordé ces objets en tant que données résultant de leur travail de récolte, d'autres les ont connus comme données à traiter par des outils dont ils avaient la maîtrise technique. Il en résulte deux points de vue, celui des données et celui de la technique, qui sont plus ou moins partagés par chacun selon son parcours et son travail. Ces deux points de vue complémentaires se font sentir aussi bien lorsqu'il s'agit des outils à employer avec ces données textuelles, que lorsqu'on aborde les questions méthodologiques relatives à ces mêmes données.

OUTILS

Dans le domaine des outils employés, l'éventail des pratiques est très large. En effet, si certains se contentent de faire des chercher/remplacer sur LibreOffice et de faire appel aux informaticiens du laboratoire pour des tâches plus complexes, d'autres en revanche ont une maîtrise et une expérience d'outils spécialisés (Iramuteq), d'outils génériques dans un domaine précis (TXM), ou encore de langages très génériques (R, Python, Perl) mais toujours avec des modules spécialisés.

La question récurrente est celle de la valeur ajoutée d'un investissement dans un outil ou un langage sans savoir à l'avance s'il répondra aux besoins, ni connaître son évolution future et sa pérennité. Il est également difficile de connaître à l'avance si cet outil va facilement s'insérer dans notre chaîne de traitements, ni si les utilisateurs se l'approprieront facilement.

L'intérêt d'outils assez simples comme Open Refine apparaît dans les tâches de mise en forme ou de nettoyage des données. Néanmoins les outils clés en main ont parfois l'inconvénient de ne pas garder une trace des traitements effectués sur les données.

Le débat qui a suivi a fait clairement émerger plusieurs classes successives dans les traitements à faire subir aux données textuelles : le nettoyage ou la mise en forme (parfois précédés ou mêlés à une tâche d'exploration), l'annotation et enfin l'analyse.

Les participants ont manifesté un grand intérêt pour la confection d'un état des lieux des outils existants, avec leur placement dans cette chaîne de traitements. Il serait intéressant de faire ce travail en coordination avec les suites de l'atelier « tableau des compétences ».

DONNÉES ET CONNAISSANCES

En ce qui concerne les données, certains participants ont clairement demandé à mieux connaître les formats de données (par ex. la TEI) ainsi que les dépôts de données textuelles dépendant de structures comme Huma-Num ou Clarin.

Et enfin, dans le domaine de la linguistique, la conscience d'un risque de biais par un traitement via un outil « boîte noire » est générale, mais la connaissance précise du mécanisme de ces biais lors des traitements n'est souvent partagé que par celles et ceux qui ont un bagage linguistique et qui peuvent situer le problème au niveau de la segmentation, de la suppression des mots-outils, ou encore de la lemmatisation. Le moyen d'éviter ces biais est, selon la majorité des participants, de pouvoir contrôler le processus de traitement, et de revenir aussi souvent que possible aux données brutes pour les vérifications nécessaires.

Un besoin de formation a émergé sur ce sujet des connaissances, et pourrait se réaliser par un Tuto ou un autre type de formation.

En conclusion, nous pouvons dire que la qualité du travail sur les données textuelles est liée à deux types d'expertise : celle des données, et celle des outils ; lorsque ces deux expertises sont partagées, la collaboration est nécessaire mais également fructueuse.