Questions juridiques et éthiques

Bonjour,

Travailler sur un ou des corpus, ça veut dire avant tout travailler sur des données authentiques, c'est-à-dire travailler sur des textes écrits ou des enregistrements qui relèvent d'une utilisation la plus écologique possible de la langue. Ce terme d'écologique ici veut dire que l'on considère le matériau recueilli comme représentatif de l'utilisation naturelle de la langue, telle qu'elle se produit en-dehors de toute situation de recherche. Quand on recueille des courriers, papiers ou électroniques, ou quand on enregistre les interactions avec des clients à la poste, le matériau langagier est produit par quelqu'un, dans une situation de communication authentique, qui parfois donne accès à la vie privée.

Il est alors indispensable de prendre très sérieusement en considération le respect et la protection de la vie privée ou du droit d'auteur.

Je voudrais donc attirer l'attention sur certains aspects juridiques, pour vous permettre d'identifier des points de vigilance et vous inciter à chercher de l'information complémentaire si vous vous trouvez dans l'un des cas de figure envisagés ici

Si on veut récupérer du texte en grande quantité et facilement, pourquoi ne pas piocher dans ce qui semble être un réservoir inépuisable, c'est-à-dire le web ? En quelques clics, pour peu que l'on maitrise quelques outils, on peut aspirer de grandes quantités de textes de toutes sortes, rapidement, avec assez peu d'efforts.

Mais attention, ce n'est pas parce qu'un texte, quel qu'il soit, est publié sur le web, qu'il est libre de droits et que l'on peut se l'approprier à sa guise. Un texte, même non littéraire, a un auteur et cet auteur reste propriétaire de son texte, bien qu'il le donne à lire à tous. Le propriétaire du texte peut aussi être le site ou le service par lequel le texte est diffusé. Prenons un exemple.

Considérons un site bien connu d'échange de recettes de cuisine : Marmiton. On pourrait penser que comme ces recettes sont écrites par « tout le monde », et bien « tout le monde » peut les prendre et en faire ce qu'il veut. Pas du tout ! Regardons les conditions d'utilisation du site, elles indiquent l'interdiction de reproduire les recettes ou d'en extraire des données autrement que pour un usage privé.

Donc imaginons que je veuille travailler sur les recettes de cuisine, par exemple pour comprendre comment sont explicitées les opérations de transformation des ingrédients, pour savoir quel vocabulaire est privilégié par des cuisiniers non professionnels. Je peux analyser les textes dans le cadre de la recherche, mais je n'ai pas le droit de mettre à disposition les recettes du site avec des enrichissements linguistiques que j'y aurais apporté, par exemple un balisage des termes de cuisine, une annotation des instructions ou autre. Si je veux faire cela, je vais devoir contacter la société Marmiton et leur demander leur autorisation.

Cela ne veut pas dire qu'il faille s'interdire de moissonner le web pour récolter des données textuelles, il est possible à un chercheur ou un laboratoire de recherche de négocier des droits pour la recherche. À l'université de Toulouse, parmi les corpus mis à disposition sur le site REDAC par le laboratoire de linguistique CLLE-ERSS, on peut télécharger un corpus nommé Géopo. Il est constitué d'articles scientifiques qui ont été aspirés sur le site web de l'Institut Français des Relations Internationales pour des analyses linguistiques. Des droits ont ensuite été accordés par l'IFRI à l'auteure du corpus, ce qui permet de mettre celui-ci à

disposition de tous avec divers enrichissements : un balisage des parties textuelles, titres, paragraphes, etc., une annotation de certains éléments sur lesquels porte la recherche.

Mais à priori, cette mise à disposition n'aurait pas été possible sans négociation, parce que ces contenus ne sont pas libres, même s'ils sont accessibles gratuitement sur Internet.

Pour recueillir facilement un corpus, on peut collecter les contenus qui sont proposés sous des licences de diffusion ouvertes, telles que les licences Creative Commons. Elles sont repérables par un logo qui aide à préciser le contenu de la licence.

L'ensemble de wikipedia est accessible sous licence Creative Common. On peut donc, si on a par exemple besoin de se constituer un corpus multilingue, télécharger des pages de wikipedia dans diverses langues, faire les recherches et remettre ces pages à disposition des autres chercheurs avec toutes les annotations que l'on veut.

En-dehors d'une telle licence, tout contenu publié sur le web est protégé par le droit d'auteur, tout comme le contenu publié par un éditeur, Le droit d'auteur varie d'un pays à l'autre, en France il couvre généralement les 70 ans qui suivent le décès de l'auteur. Pour les textes protégés, on peut citer de courts extraits de textes, mais pas redistribuer la totalité d'un texte qui n'est pas libre. Il y a toutefois de nombreux textes qui font partie du domaine public : les textes anciens, les discours présidentiels, par exemple. Mais en dehors de ceux-là, il faut être très prudent : malgré la facilité d'accès, tout ce qui est sur le web ne peut pas être utilisé librement.

C'est ennuyeux de ne pas travailler sur un matériau libre, parce que quand on fait de la recherche, il est important que les autres chercheurs auxquels on expose nos conclusions puissent accéder à nos données et mener aussi leurs propres analyses sur les mêmes données. Il est donc préférable de privilégier des corpus non soumis à des restrictions de droits.

Une autre voie est celle du recueil de données auprès d'une population définie ou de l'ensemble de la population. Je prendrai là deux exemples pour illustrer les questions qui se posent et qui sont surtout des questions d'éthique et de protection de la vie privée.

Peut-être avez-vous entendu parler des divers travaux de recherche autour des nouvelles formes de communication. Un grand projet de recherche nommé sms4science a lancé dans plusieurs pays un appel pour que chacun donne ses sms à la science, afin de constituer un grand corpus de sms, au sein duquel on pourrait observer les transformations, les adaptations de la langue pour ce format de message court, les régularités, l'expression des émotions, les styles selon l'âge et le profil sociologique des auteurs, etc.

Là à priori pas de droit d'auteur, mais ça ne veut pas dire qu'il n'y avait pas de questions à résoudre avant la mise à disposition des données. D'abord on ne devait donner que ses propres sms, ceux dont on est l'auteur, pas ceux que l'on a reçu, c'est cohérent avec l'idée que l'on a des droits sur ce que l'on a produit. Ensuite on donnait une autorisation d'utiliser les sms envoyés dans le projet de recherche. En contrepartie, le laboratoire de recherche s'engage au respect de la vie privée en ne diffusant bien sûr ni numéro de téléphone, ni adresse mail ou autre et en anonymisant les sms recueillis. Ce qui signifie remplacer les noms de personnes, les noms de lieu, enfin tout ce qui pourrait permettre d'identifier qui que ce soit. De plus, si on recueille des données personnalisées telles que l'âge, le sexe, la profession, le niveau d'études, etc. des donneurs de sms, il est nécessaire de déclarer à la

CNIL avant la collecte du corpus la nature des données conservées et la façon dont on les conserve, dont on les exploite et dont on les sécurise pour éviter qu'elles soient piratées.

Le même genre de questions se pose lorsque l'on collecte des textes scolaires ou universitaires, comme nous l'avons fait au LIDILEM avec les corpus Littéracie Avancée ou Scoledit. Ce sera mon 2e exemple. Si vous acceptez de donner par exemple votre mémoire de master, ou un dossier que vous avez rédigé dans le cadre de vos études, ou si vous êtes un collégien ou lycéen, donc mineur, et que vous donnez ce que vous avez fait pour une évaluation, il est du devoir du chercheur de vous protéger, de faire en sorte que l'on ne puisse pas faire un lien entre vous et tel ou tel document. Personne en-dehors de vous n'a à savoir comment vous écrivez ou écriviez à 8 ans, 15 ans ou 25 ans, si votre orthographe est bonne ou mauvaise, si vous savez tourner de jolies phrases, si vous avez de bonnes idées. Le chercheur se doit d'informer précisément toute personne qui accepte de donner un écrit qu'elle a produit de ce à quoi cet écrit va servir et se doit de garantir son anonymat.

Ceci pose de vrais problèmes dès lors que le corpus est constitué d'enregistrements audios ou vidéos. Donc pour terminer, voyons comment un corpus d'oral qui s'appelle CLAPI a géré ces questions :

Les participants sont sollicités pour donner leur consentement éclairé à l'utilisation des données. Un consentement éclairé signifie de comprendre comment les données vont être utilisées. Le chercheur doit donc être le plus précis possible dans son explicitation de la recherche prévue.

L'objectif est ensuite de préserver autant que possible l'anonymat et la vie privée. Ce qui implique pour les chercheurs une réflexion très poussée sur ce qui permet d'identifier quelqu'un. On pourrait imaginer qu'il suffit de remplacer dans les transcriptions les noms de personne, par exemple XXX au lieu de Fernand Dupont. Mais ce n'est pas si simple. Si on dit par exemple, « mon cousin qui habitait Impasse Jean-Jaurès en 1988 », que un peu auparavant on a donné l'âge du cousin, et que dans l'Impasse Jean-Jaurès en 1988, il y a une seule famille avec un garçon de cet âge, le cousin peut être identifié. Donc il faut aussi prêter attention aux noms de lieu, aux dates, etc. Toutefois, les données doivent continuer à être exploitables pour la recherche. On ne peut donc pas effacer purement et simplement les passages dans lesquels il y a un nom propre, de personne ou de ville, le mieux est de les remplacer par une expression équivalente.

De la même manière, dans les enregistrements audio, on peut déformer le signal de façon à rendre la voix méconnaissable, ou bien remplacer une expression par un beep ou un silence…

Pour les enregistrements vidéo, c'est un peu plus compliqué bien sûr, il est difficile d'anonymiser. On pourrait flouter les visages et déformer la voix, mais si la recherche porte par exemple sur les mimiques accompagnant l'interaction, on doit pouvoir analyser les visages. Le consentement éclairé est donc la garantie minimale pour les participants.

En aucun cas, on ne doit enregistrer des gens à leur insu et transformer les enregistrements en corpus publics.

En conclusion, chaque constitution de corpus doit penser bien en amont les moyens de collecter du matériau linguistique dans le respect du droit d'auteur, de la vie privée et en cohérence avec l'éthique de la recherche.