Charte Éthique et Big Data : parce que mon corpus le vaut bien !

Alain Couillault [1], Karën Fort [2]
Université de La Rochelle/L3I, Université de Lorraine/LORIA

Mots-clés : corpus, éthique, ressources langagières, documentation.

1. Motivations

Les corpus, en particulier annotés, sont aujourd'hui indispensables non seulement pour les études linguistiques, mais également pour la création et l'évaluation d'outils de traitement automatique.

La constitution et l'annotation de ces corpus mobilisent des ressources importantes et leur financement est souvent assuré, au moins en partie, par des projets de type ANR. Or, le Web 2 a permis l'apparition de plate-formes de myriadisation du travail parcellisé (*microworking crowdsourcing*), dont `Amazon Mechanical Turk`, qui proposent à des demandeurs (*Requesters*) d'accéder à une « foule » de travailleurs (*Turkers*), qui sont très peu, voire pas du tout, rémunérés. Le succès de ce type de plate-forme dans le domaine du traitement automatique des langues pose non seulement des problèmes éthiques (les *Turkers* ne bénéficient pratiquement d'aucun droit), mais également de qualité des données produites et de la propriété intellectuelle de celles-ci (Sagot *et al.* 2011). Surtout, ces systèmes se développant, le coût à la `Amazon Mechanical Turk` pourrait devenir une norme de fait pour les agences de moyens, empêchant ainsi le développement de corpus ré-utilisables tels que nous les connaissons aujourd'hui. Cette situation pourrait être contre-productive, en orientant les financements vers des projets présentant un risque juridique, dont la pérennité et la qualité des données seraient mal maitrisées, et qui favoriseraient des sous-emplois à bas revenus, hors du territoire.

Par ailleurs, si de gros efforts ont été faits pour sensibiliser les producteurs de corpus à l'adoption de bonnes pratiques, que ce soit en matière d'encodage et de création (Wynne 2005), notamment à travers les formats TEI [1] ou Dublin Core [2], ou d'éthique et de législation (Baude *et al.* 2006), trop de corpus sont encore trop peu documentés.

La Charte Éthique et Big Data vise à répondre à ces deux préoccupations : une fois adoptée par les agences de moyens, elle permettra d'encourager les bonnes pratiques, à la fois en termes de documentation et de respect de la législation, et de respect des droits des personnes travaillant sur la ressource. Elle est auto-déclarée et se présente sous la forme d'un document à remplir par

1. Université de La Rochelle / Laboratoire L3I, alain.couillault@univ-la-rochelle.fr
2. Université de Lorraine / LORIA, karen.fort@loria.fr

1. `http://www.tei-c.org/index.xml`
2. `http://dublincore.org/`

les producteurs de ressources. Le document lui-même est neutre et n'encourage, ni ne décourage, les comportements non éthiques. Il liste cependant des informations à fournir et, par là, impose une documentation minimale de la ressource. Nous pensons que son adoption par les agences de moyens devrait favoriser la diffusion de bonnes pratiques en matières d'éthique, de propriété intellectuelle, de traçabilité et de qualité.

2. Démarche

2.1. De l'éthique et du « Big Data » ?

La Charte est « éthique » en ce qu'elle promeut les bonnes pratiques en matière de données personnelles, de respect du droit du travail, de propriété intellectuelle de tous les acteurs, et, en outre, instaure un contrat de confiance entre les fournisseurs et les utilisateurs. Le terme de « Big Data » recouvre une grande variété de données, qui sont volumineuses ou complexes. La disponibilité en grand volumes de données fait apparaître des problématiques nouvelles, ou exacerbes des problématiques existantes (comment connaître la propriété intellectuelle d'un jeu de données volumineux et de sources variées, comment garantir la « véracité » ou la qualité des données, comment assurer la mise à jour de données changeantes ...). Les corpus (oraux ou écrits) s'inscrivent dans cette perspective d'ensemble et bénéficient de la charte. Celle-ci est également adaptée à des ressources langagières de taille modeste (les lexiques, par exemple), lorsqu'elles présentent une certaine complexité.

2.2. Processus de rédaction

La rédaction de cette charte est le fruit d'une collaboration entre : (i) des associations, APROGED (Association des professionnels du numérique), ATALA (Association pour le Traitement automatique des LAngues), AFPC (Association Française de Communication Parlée), (ii) des acteurs industriels (Eptica/Lingway, Digital Ethics) et (iii) des organismes de financement qui ont soit contribué directement, soit apporté leur soutien au projet. ELDA/ELRA et le CERSA (Centre d'Etude et de Recherche de Sciences Administratives et politiques) ont également participé aux réunions. Ce travail s'est déroulé de juin à décembre 2012, à raison d'une réunion par mois, chaque rédacteur contribuant à un wiki dont le contenu était revu lors des réunions.

3. Contenu

La charte est proposée sous la forme d'un formulaire destiné à accompagner le corpus ou le dossier de demande de financement. Elle comporte trois volets majeurs qui couvrent la traçabilité, la propriété intellectuelle et les législations spécifiques, complétés par une section dédiée à la description du corpus. Chacun de ces volets est organisé en trois sections couvrant l'amont du travail de constitution ou de transformation du corpus (*avant*), les considérations liées au travail de constitution ou de transformation lui-même (*pendant*), et les aspects concernant une éventuelle distribution du corpus (*après*).

3.1. Éthique

La Charte Éthique et Big Data impose de documenter les moyens mis en oeuvre pour la création du jeu de données, notamment de préciser le statut et le mode de rémunération des personnes ayant participé ou devant participer au projet. En cas d'utilisation d'une plate-forme de myriadisation, il est demandé de la spécifier, d'indiquer les critères de sélection des travailleurs et le mode et le montant de la rémunération. L'utilisateur éventuel de données et le financeur peuvent ainsi choisir une ressource plutôt qu'une autre en fonction de leur propres critères.

3.2. Aspects juridiques : propriété intellectuelle et législation spécifique

Bien qu'il existe de nombreuses licences génériques, les aspects liés à la propriété intellectuelle sont trop souvent un frein à la constitution, l'utilisation ou la distribution de corpus : licence non-documentée, licence trop restrictive, manque de clarté des Conditions Générales d'Utilisation des sites Internet, difficulté à identifier le propriétaire des données. Le risque est alors de se mettre hors la loi ou de ne pas pouvoir diffuser son corpus.

Par ailleurs, les questions liées aux législations spécifiques nécessitent une attention particulière pour des corpus linguistiques oraux ou écrits, lorsque l'on collecte des données produites par des personnes physiques (extraits de textes, pages Web, enregistrements…). Il est en effet nécessaire de s'assurer que les différentes législations concernant les droits des personnes (propriété intellectuelle, respect de la vie privé) sont respectées, notamment lorsqu'elles imposent d'obtenir l'autorisation d'usage ou la cession de droits de la part de celles-ci.

3.3. Traçabilité

Parce qu'un corpus n'apparaît pas *ex-nihilo*, il faut pouvoir retracer son historique, au-delà de son versionnement. La charte prend en compte différentes situations, selon que les données sont créées par le fournisseur lui-même (données primaires) ou par agglomération de données existantes, avec ou sans transformation de la part du fournisseur. Dans ce dernier cas, la traçabilité est facilitée si une Charte Éthique et Big Data est déjà disponible pour les données utilisées en amont. Dans le cas contraire, il est nécessaire de fournir l'ensemble des renseignements permettant de retracer l'origine des ressources et d'assurer ainsi la conformité des données fournies aux différentes exigences (qualité, propriété intellectuelle, licence, etc.).

La Charte Éthique et Big Data impose de décrire le processus de transformation, ainsi que les outils et intervenants impliqués dans ce processus.

3.4. Qualité

La Charte Éthique et Big Data permet de préciser les moyens mis en œuvre pour assurer la qualité de la ressource, mais elle ne les impose pas. L'utilisateur peut ainsi choisir d'utiliser les corpus quelle que soit leur qualité et peut éventuellement mettre en place son propre processus d'assurance qualité en toute connaissance de cause.

4. Exemple de charte : le corpus TCOF-POS

Nous avons obtenu l'accord de C. Benzitoun pour rédiger la Charte Éthique et Big Data correspondant au corpus TCOF-POS. Ce corpus présente l'avantage d'avoir été réalisé relativement récemment et d'être bien documenté, à travers son site Web [3] et l'article le décrivant (Benzitoun *et al.* 2012). Il nous a fallu environ une heure et demie pour remplir la charte [4], que nous avons ensuite fait réviser par C. Benzitoun. Nous avons constaté que si celle-ci est relativement concise, elle n'est pas toujours assez explicite. Une documentation, notamment par le biais d'exemples bien choisis, permettrait non seulement de faciliter son utilisation, mais également d'apporter une dimension pédagogique à la charte, ainsi que des propositions de solutions, notamment en matière de choix de licence.

5. Conclusion et perspectives

La Charte Éthique et Big Data se veut un outil de pérénisation des corpus, encourageant la qualité des ressources créées, leur diffusion et le respect des personnes travaillant à la création de celles-ci. Cet outil est perfectible, notamment en ce qui concerne sa documentation, que nous améliorons actuellement. À ce jour (mars 2013), la Charte a été adoptée par Cap Digital (co-rédacteur). Nous avons également proposé à la DGLFLF, l'ANR et la DGA de l'utiliser. La Charte Éthique et Big Data est disponible sous plusieurs formes : un wiki [5], un fichier pdf et une version en anglais.

Bibliographie

Baude O., Blanche-Benveniste C., Calas M.-F., Cappeau P., Cordereix P., Goury L., Jacobson M., De Lamberterie I., Marchello-Nizia C. & Mondada L. (2006), *Corpus oraux, guide des bonnes pratiques 2006*, CNRS Editions, Presses Universitaires Orléans.

Benzitoun C., Fort K. & Sagot B. (2012), « TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe », in : *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France : 99–112.

Sagot B., Fort K., Adda G., Mariani J. & Lang B. (2011), « Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé », in : *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France, 12 pages.

WYNNE M. (Ed) (2005), *Developing Linguistic Corpora : a Guide to Good Practice*, Oxford : Oxbow Books.

3. http://cnrtl.fr/corpus/perceo/

4. Voir : http://wiki.ethique-big-data.org/chartes/CharteEthiqueBigDataLightTCOFPOS.pdf.

5. http://wiki.ethique-big-data.org