

# Analyse économique, juridique et éthique du crowdsourcing pour le TAL

Gilles Adda\*, Joseph J. Mariani\*, Laurent Besacier †

\*LIMSI-CNRS & IMMI-CNRS † LIG/UJF

## Abstract

L'utilisation du crowdsourcing et en particulier de la myriadisation du travail parcellisé est devenu un outil largement utilisé dans les communautés utilisant les données et leur annotation, en particulier avec l'avènement du "Big Data". Ceci est particulièrement vrai pour la communauté parole, où les tâches mises en œuvre (par exemple transcription, traduction) sont bien adaptées à l'utilisation de cette technique. Suivant de nombreux articles questionnant l'éthique du crowdsourcing rémunéré, nous mettons en avant certaines bonnes pratiques à mettre en œuvre lors de l'utilisation du crowdsourcing, fondées sur une analyse économique, légale et éthique.

## 1. Introduction

Au tournant des années 90, le domaine du TAL, et en particulier le domaine du traitement de la parole a changé drastiquement, avec le développement de méthodes d'apprentissage probabiliste et l'établissement du paradigme de l'évaluation. Un résultat important de ces changements est le développement nécessaire de corpus de taille toujours plus importante et de toute nature. De corpus de parole de quelques heures et ou de corpus de texte de quelques millions de mots, les modèles que nous développons aujourd'hui sont fondés sur des données issues de multiples médias, modes et langues (3M). Nous traitons dans les systèmes état-de-l'art des milliers d'heures de parole et de vidéo, des milliards de mots de texte de toute nature (journaux, web, médias sociaux, sous-titres, transcription, traductions, etc.).

La constitution de ces données a reposé au départ sur le travail artisanal des chercheurs (et de leurs étudiants) et d'agences spécifiques qui ont été chargées également de distribuer ces corpus, comme le LDC<sup>1</sup> ou ELDA<sup>2</sup>. Mais, en suivant les fameux 3V du "Big data" (Volume, Vitesse et Variété) la taille des corpus nécessaires rend les méthodes classiques de recueil et d'annotation obsolètes. Celles-ci sont fondées le plus souvent sur un choix précis, dans un cadre juridique défini, et des processus d'annotation utilisant soit des volontaires (chercheurs, étudiants) soit des salariés, et où les problèmes d'éthique sont traités dès le départ de la conception du corpus ; ces méthodes, qui assurent dans une certaine mesure la qualité et l'intégrité des données, sont en effet coûteuses en temps et en argent. Ainsi, pour traiter les corpus dont les 3V sont de taille croissante, dans un temps et dans un budget contraint, il a fallu mettre en place de nouvelles méthodes, et en particulier le *crowdsourcing*, sous sa forme de microworking, aussi dénommé myriadisation du travail parcellisé (MTP) (Sagot et al., 2011). Dans cet article, nous utiliserons comme terminologie :

**crowdsourcing** l'ensemble des services de crowdsourcing, tel que décrit dans la section 2. ;

MTP les services de microworking rémunéré ;

**service** un système spécifique de MTP ;

**fournisseur** la personne ayant un tâche à effectuer ;

**travailleur** la personne rétribuée pour effectuer une tâche.

De nombreux auteurs, dont nous-mêmes, ont mis en garde sur les problèmes éthiques et pratiques que l'utilisation du crowdsourcing amène (Adda and Mariani, 2010; Fort et al., 2011; Sagot et al., 2011; Adda et al., 2013; Fort et al., 2014a).

Un résultat pratique de ces mises en garde a été la constitution de la *Charte Éthique et Big Data* (Couillaud and Fort, 2013). L'objectif de cette Charte est de fournir aux développeurs des ressources, non seulement dans la communauté du TAL, mais plus largement dans la communauté Big Data, un cadre précis pour documenter leurs ressources et assurer leur traçabilité et leur transparence. La charte permet de mettre en œuvre un certain nombre de bonnes pratiques.

Cependant, l'analyse économique et légale du *crowdsourcing*, détaillée en particulier dans (Adda et al., 2013), nous conduit à proposer un certain nombre de recommandations spécifiques, propres à limiter les risques en particulier éthiques.

## 2. Crowdsourcing et myriadisation du travail parcellisé

*Crowdsourcing* est un terme fourre-tout, qui regroupe toutes sortes de systèmes, allant de WIKIPÉDIA<sup>3</sup> à AMAZON MECHANICAL TURK<sup>4</sup>, en passant par des projets de science collaborative comme FOLDIT<sup>5</sup> ou des plateformes de travail artistique comme 99DESIGNS<sup>6</sup>. Une visualisation (partielle!) du paysage du crowdsourcing, avec une optique économique a été faite par Dawson&Byngall (voir figure 1). Nous nous intéressons ici plus particulièrement au *Microworking* rémunéré, ou MTP<sup>7</sup> dont MTurk est le représentant le plus connu, car il est à la fois la forme

<sup>3</sup>[www.wikipedia.org](http://www.wikipedia.org)

<sup>4</sup>[www.mturk.com/mturk](http://www.mturk.com/mturk)

<sup>5</sup>[fold.it](http://fold.it)

<sup>6</sup>[99designs.com](http://99designs.com)

<sup>7</sup>myriadisation du travail parcellisé

<sup>1</sup>Linguistic Data Consortium, [www ldc.upenn.edu/](http://www ldc.upenn.edu/)

<sup>2</sup>Evaluations and Language resources Distribution Agency, [www.elda.org/](http://www.elda.org/)

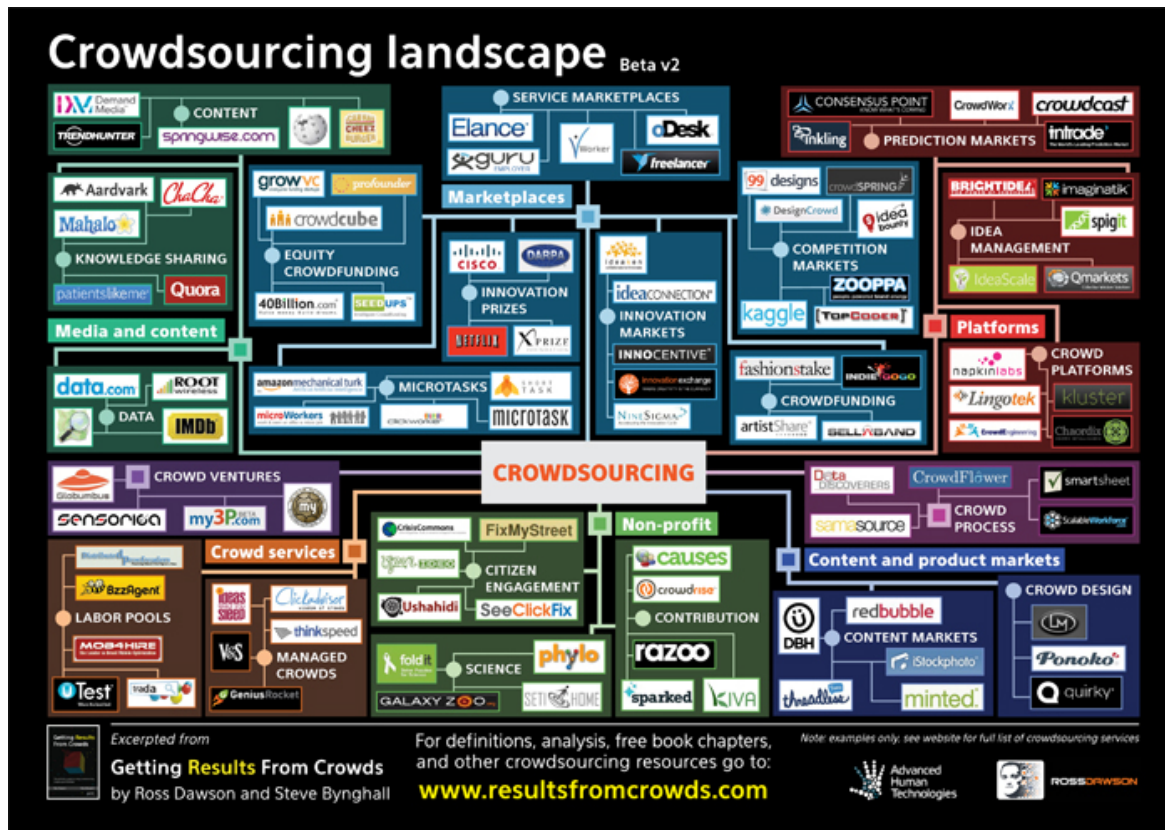


Figure 1: Paysage du Crowdsourcing, d'après Dawson&Byngghall

de crowdsourcing la plus utilisée en TAL au niveau international, et celle qui pose potentiellement le plus de problèmes.

### 3. Problèmes économiques et juridiques

Les principaux problèmes viennent de la relation du travailleur impliqué dans le MTP avec le service de *crowdsourcing*, le fournisseur de la tâche, et la tâche elle-même. Il y a un problème éthique ou juridique lorsque la tâche est assimilable à une expérience sur des humains ou si elle est assimilable à un vrai travail au sens légal du terme. Dans les deux cas, les chercheurs en TAL ayant recours au MTP ont des obligations éthiques. La tâche pourrait être assimilable à une expérience sur des humains dans le cas où on recueille une propriété appartenant en propre aux travailleurs (par exemple leur parole ou leurs opinions). Dans ce cas, des règles existent concernant le paiement de sujets humains soumis à une expérience. En particulier il faut s'assurer que la rémunération ne constitue pas une coercition ; il convient également de faire signer un accord de consentement éclairé, expliquant en particulier quel(s) usage(s) sera fait de la propriété recueillie. Il faut cependant relever que la jurisprudence ne considère pas que par exemple le recueil de parole relève d'expériences sur des sujets humains. Nous ne détaillerons donc pas plus avant cette voie.

La tâche peut correspondre à un travail, lorsque le revenu généré par l'activité représente une partie significative du

revenu du travailleur, et/ou si la tâche *en soi* est considéré hors du champ du crowdsourcing, comme un travail. En particulier, pour des tâches comme la transcription de parole, ou la traduction, les chercheurs ont fait appel depuis des décennies à des agences spécialisées comme LDC ou ELDA employant des salariés ; pour ces tâches qui sont d'habitude effectuées par des travailleurs salariés, le MTP est assimilable à une délocalisation sur le web. Il faut souligner que d'un point de vue légal, les travailleurs en MTP sont des travailleurs indépendants, et donc ne relèvent pas des législations du travail concernant le revenu minimum ou la durée du travail.

Parmi l'ensemble des paramètres économiques et légaux à prendre en compte (Adda et al., 2013), nous en relevons trois cruciaux, le salaire, l'anonymat et la législation du travail.

#### 3.1. Le salaire

Un problème récurrent est la faiblesse des rétributions : en dessous de 2\$/heure (Ross et al., 2009; Ipeirotis, 2010; Chilton et al., 2010) dans MTurk. Ce taux est significativement plus bas que les taux légaux (7,25\$ aux USA, 6,50£ (~10,53\$) au Royaume-Uni, 9,53€ (~12,00\$) en France), ce qui implique des problèmes éthiques mais également économiques et légaux, de concurrence faussée. De plus, il arrive fréquemment que les fournisseurs payent tard, voire pas du tout (Silberman et al., 2010). Concernant le salaire,

nous pouvons souligner avec (?), ‘Pay enough or not pay at all’ ; en effet, seules deux solutions semblent à la fois éthiques et converger vers un système efficace et durable :

- donner une rétribution substantielle, et modulée en fonction de la qualité observée,
- ne pas donner de compensation monétaire, mais faire appel au volontariat ; cette solution, comme par exemple celle présentée dans (Fort et al., 2014b), convient bien à de nombreuses applications en TAL qui manquent de moyens pour appliquer la première solution !

### 3.2. L’anonymat

Une caractéristique importante des services actuels de crowdsourcing engendre un faisceau de problèmes : l’anonymat. L’anonymat est censé protéger à la fois le fournisseur et le travailleur de contraintes ciblées et de spams au sens large ; il engendre en fait une rupture de relations entre les deux partenaires, qui conduit par exemple au rejet de tâches effectuées sans possibilité de recours, l’impossibilité de se regrouper entre travailleurs d’un même fournisseur ou d’un même service. Elle peut permettre également de cacher la nature du travail à effectuer, la tâche elle-même n’étant qu’un élément d’un ensemble, et donc conduire à des dérives (Zittrain, 2008).

### 3.3. La législation

En suivant la législation en vigueur, il est difficile de décider si les travailleurs du MTP devraient être considérés comme des salariés ou non (Adda et al., 2013; Felstiner, 2011; Wolfson and Lease, 2011). Pour le moment, le travail sur internet, en particulier le MTP, est dans une “zone grise”, où la régulation actuelle n’est pas pertinente. Mais il est probable que si le marché du MTP s’accroît, les législations nationales et internationales devront revoir les lois du travail pour réguler le marché du travail sur internet (Felstiner, 2011; Wolfson and Lease, 2011). Cette régulation est nécessaire à moyen terme, à la fois pour protéger le travailleur, mais également pour stabiliser économiquement le marché, et assurer un cadre légal clair aux fournisseurs de tâches.

## 4. Recommandations

Un rappel des problèmes est toujours utile, mais vain s’il n’existe pas de solutions. Tel n’est pas le cas des problèmes éthiques soulevés par l’emploi des MTP. Nous donnons ici une série de suggestions issues de différents articles (Wolfson and Lease, 2011; Adda et al., 2013; Chiarella, 2011; Bederson and Quinn, 2011) en distinguant les solutions relevant d’une “bonne pratique” lors de l’utilisation du *crowdsourcing* en TAL, et de ce qui relève d’une politique à moyen et long terme.

### 4.1. Bonnes pratiques en MTP

Nous donnons ici une liste de recommandations, compilée d’après différents articles et études, et pouvant permettre de limiter tant les problèmes éthiques que les problèmes de qualité, lors de l’emploi des MTP :

- (i) Payez bien : Il ne sert à rien de chercher à sous-payer les travailleurs ; outre le risque d’exploitation, vous risquez de diminuer la qualité et de ne pas trouver de volontaires.
- (ii) Une rémunération horaire : le salaire doit dépendre du temps nécessaire pour effectuer la tâche.
- (iii) Annoncez le salaire : affichez le salaire horaire estimé.
- (iv) Économisez le temps : optimisez les tâches, afin que les travailleurs utilisent leur temps uniquement pour ce qui est utile.
- (v) Des mesures de qualités objectives : la décision de rejeter ou non un travail doit être fondée sur des mesures objectives, définies à l’avance et accessibles au travailleur.
- (vi) Donnez un feedback immédiat sur la qualité du travail : donnez le feedback dès que possible, en utilisant les métriques définies.
- (vii) Donnez un feedback sur le travail global : si le travail fourni pose problèmes, donnez des avertissements au travailleur à l’avance.
- (viii) Annoncez l’échéance du paiement.
- (ix) Suivez les règles de paiement : Payez dès que possible, et toujours avant le terme défini à l’avance.
- (x) Donnez le contexte de la tâche : Certaines tâches pouvant conduire le travailleur à participer à quelque chose qu’il réproouve, il faut toujours donner objectivement le contexte dans lequel le travail demandé s’inscrit.
- (xi) Soyez respectueux des lois : Les fournisseurs de tâches doivent s’inspirer des lois en vigueur dans leur pays, afin de protéger les travailleurs du MTP et préserver l’emploi local.
- (xii) Utilisez des contrats pour définir clairement vos relations avec les travailleurs : La relation fondée uniquement l’acceptation de termes par un clic souris n’est pas suffisante. Définissez un contrat clair et concis, et assurez vous qu’il est compris par le travailleur, cela limitera la plupart des problèmes potentiels.
- (xiii) Soyez ouvert et honnête : En énonçant à l’avance ce que vous attendez précisément des travailleurs, cela vous permettra d’adapter la qualité de leur travail à cette attente.
- (xiv) Limitez l’anonymat : l’anonymat des fournisseurs conduit à pouvoir rejeter un travail correct en toute impunité. Cela permet aussi de proposer des tâches non éthiques voire illégales. L’anonymat des travailleurs induit un phénomène de tricherie généralisé sans risques.
- (xv) Donnez une possibilité de requête : Il est important de prévoir de donner au travailleur une explication précise si un travail a été rejeté, et de pouvoir corriger le rejet, en cas d’erreur.

## 4.2. Développement des MTP en TAL

Au-delà des bonnes pratiques, nous pouvons donner quelques pistes pouvant à terme permettre de développer des services MTP utilisables en limitant structurellement les problèmes éthiques :

- (i) Développer une plateforme MTP comme un service d'intérêt public au profit de la communauté de la recherche et de l'enseignement, permettant de produire des corpus et des annotations. On peut imaginer plusieurs sortes de plateformes :
  - une plateforme à destination d'annotateurs spécialistes rémunérés, et assurant l'éthique, le droit et la qualité (application du droit du travail, d'un minimum garanti, pas d'anonymat, . . .) ;
  - une plateforme à destination des non-spécialistes volontaires, pour les tâches où il est nécessaire d'avoir accès à une fraction représentative de la population ;
  - une plateforme de jeux sérieux, comprenant une aide au développement de jeux et un accès à une communauté de joueurs.
- (ii) Constituer et maintenir une communauté mondiale de spécialistes des différentes langues à laquelle les chercheurs pourront faire appel.
- (iii) Développer dans l'ensemble des pays, un réseau de contacts permettant d'accéder aux locuteurs des langues en danger, et qui pourraient assurer à la fois une rémunération conforme aux us et coutumes locaux, un traitement éthique des locuteurs et des données recueillies, et assurer l'annotation des données.
- (iv) Développer des tests (en particulier de connaissance d'une langue ou d'une variété de langue) permettant de préqualifier les travailleurs :
  - établir les moyens de conduire des tests fiables, en utilisant des références validées par des spécialistes ;
  - établir un service indépendant permettant d'établir la réputation des travailleurs et des fournisseurs.

## 5. References

Adda, G. and Mariani, J. (2010). Language resources and amazon mechanical turk: legal, ethical and other issues. In *LISLR 2010, "Legal Issues for Sharing Language Resources workshop"*, LREC 2010, Valletta, Malta, May.

Adda, G., Mariani, J., Besacier, L., and Gelas, H., (2013). *Crowdsourcing for Speech Processing*, chapter Economic and Ethical Background of Crowdsourcing for Speech, pages 303–334. Wiley.

Bederson, B. B. and Quinn, A. J. (2011). Web workers unite! addressing challenges of online laborers. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 97–106. ACM.

Chiarella, S. (2011). Cooking with Sharon. <http://mechanicalturk.typepad.com/blog/2011/07/cooking-with-sharon-tip-3\manage-your-reputation.html>.

Chilton, L. B., Horton, J. J., Miller, R. C., and Azenkot, S. (2010). Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 1–9.

Couillault, A. and Fort, K. (2013). Charte Éthique et Big Data : parce que mon corpus le vaut bien ! In *Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasburg, France, July. 4 pages.

Felstiner, A. (2011). Working the Crowd: Employment and Labor Law in the Crowdsourcing Industry. *Berkeley Journal of Employment and Labor Law*, 32(1), August.

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2).

Fort, K., Adda, G., Sagot, B., Mariani, J., and Couillault, A., (2014a). *HLT Challenges for Computer Science and Linguistics*, chapter Crowdsourcing for Language Resource Development: Criticisms about Amazon Mechanical Turk Overpowering Use. Springer.

Fort, K., Guillaume, B., and Chastant, H. (2014b). Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 2–6, New York, NY, USA. ACM.

Ipeirotis, P. (2010). Demographics of mechanical turk. CeDER Working Papers, <http://hdl.handle.net/2451/29585>, March. CeDER-10-01.

Ross, J., Zaldivar, A., Irani, L., and Tomlinson, B. (2009). Who are the turkers? worker demographics in amazon mechanical turk. Social Code Report 2009-01, <http://www.ics.uci.edu/~jwross/pubs/SocialCode-2009-01.pdf>.

Sagot, B., Fort, K., Adda, G., Mariani, J., and Lang, B. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. In *TALN'2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France.

Silberman, M. S., Ross, J., Irani, L., and Tomlinson, B. (2010). Sellers' problems in human computation markets. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 18–21.

Wolfson, S. M. and Lease, M. (2011). Look before you leap: Legal pitfalls of crowdsourcing. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10.

Zittrain, J. (2008). *The Future of the Internet—And How to Stop It*. Yale University Press, New Haven, CT, USA.