Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé

Benoît Sagot, Karën Fort, Gilles Adda, Joseph Mariani, Bernard Lang

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé

Benoît Sagot[1] Karën Fort[2,3] Gilles Adda[4] Joseph Mariani[4,5] Bernard Lang[6]
(1) Alpage, INRIA Paris–Rocquencourt & Université Paris 7,
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
(2) INIST-CNRS, 2 allée de Brabois, 54500 Vandoeuvre-lès-Nancy, France
(3) LIPN, Université Paris Nord, 99 av J-B Clément, 93430 Villetaneuse, France
(4) LIMSI-CNRS, Bât. 508, rue John von Neumann, Université Paris-Sud BP 133, 91403 Orsay Cedex, France
(5) IMMI-CNRS, Bât. 508, rue John von Neumann, Université Paris-Sud BP 133, 91403 Orsay Cedex, France
(6) INRIA Paris–Rocquencourt, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{benoit.sagot, bernard.lang}@inria.fr, karen.fort@inist.fr, {gilles.adda,joseph.mariani}@limsi.fr

Résumé. Cet article est une prise de position concernant les plate-formes de type Amazon Mechanical Turk, dont l'utilisation est en plein essor depuis quelques années dans le traitement automatique des langues. Ces plate-formes de travail en ligne permettent, selon le discours qui prévaut dans les articles du domaine, de faire développer toutes sortes de ressources linguistiques de qualité, pour un prix imbattable et en un temps très réduit, par des gens pour qui il s'agit d'un passe-temps. Nous allons ici démontrer que la situation est loin d'être aussi idéale, que ce soit sur le plan de la qualité, du prix, du statut des travailleurs ou de l'éthique. Nous rappellerons ensuite les solutions alternatives déjà existantes ou proposées. Notre but est ici double : informer les chercheurs, afin qu'ils fassent leur choix en toute connaissance de cause, et proposer des solutions pratiques et organisationnelles pour améliorer le développement de nouvelles ressources linguistiques en limitant les risques de dérives éthiques et légales, sans que cela se fasse au prix de leur coût ou de leur qualité.

Abstract. This article is a position paper concerning Amazon Mechanical Turk-like systems, the use of which has been steadily growing in natural language processing in the past few years. According to the mainstream opinion expressed in the articles of the domain, these online working platforms allow to develop very quickly all sorts of quality language resources, for a very low price, by people doing that as a hobby. We shall demonstrate here that the situation is far from being that ideal, be it from the point of view of quality, price, workers' status or ethics. We shall then bring back to mind already existing or proposed alternatives. Our goal here is twofold : to inform researchers, so that they can make their own choices with all the elements of the reflection in mind, and propose practical and organizational solutions in order to improve new language resources development, while limiting the risks of ethical and legal issues without letting go price or quality.

Mots-clés : Amazon Mechanical Turk, ressources linguistiques.

Keywords: Amazon Mechanical Turk, language resources.

1 Introduction

Le traitement des langues a grandement évolué au cours des ces vingt dernières années, tant dans le traitement de l'écrit que de la parole. Stimulé par le paradigme de l'évaluation, le rôle des ressources linguistiques dans ce développement a été et reste crucial : elles sont à la fois matière première, objet d'étude et ressource pour l'évaluation de systèmes. Nous proposons ici une critique [1] d'un outil nouveau de constitution de ces ressources, le *microworking* par le biais du *crowdsourcing*. *Microworking* fait référence au fait que le travail est segmenté en petites tâches, *crowdsourcing* au fait que le travail est délocalisé (*outsourced*) et est effectué par un grand nombre de personnes (*crowd*), payées ou non. Nous néologiserons crowdsourcing en « myriadisation » et microworking en « travail parcellisé », et la conjonction des deux par « myriadisation du travail parcellisé ».

Nous aborderons en détails le cas d'un système de myriadisation du travail parcellisé (m.t.p. dans la suite) qui a fait florès ces derniers temps, Amazon Mechanical Turk (MTurk), notamment pour sa capacité à produire des corpus annotés à un coût très faible. Les auteurs de cet article ont contribué, à des degrés divers, à la mise en

1. Au sens d'un examen raisonné, objectif, qui s'attache à relever les qualités et les défauts et donne lieu à un jugement de valeur.

place du paradigme de l'évaluation et au développement de nombreux outils et ressources dans le domaine du traitement du langage. Nous sommes à ce titre conscients de l'importance du développement et de la diffusion de celles-ci et du frein que représente leur coût, souvent rédhibitoire. Cependant, nous voulons mettre en avant le fait que le coût du développement est un argument non fondé en ce qui concerne la m.t.p., tout d'abord parce qu'il masque des problèmes économiques complexes, ensuite parce qu'il met sous le boisseau le problème de la qualité des ressources ainsi obtenues, enfin parce qu'il omet la question de l'éthique et du droit du travail. Nous aborderons ici l'ensemble de ces questions, sans pour autant remettre en cause l'utilité de la m.t.p., à condition que son fonctionnement et son utilisation se fassent selon certains principes.

2 Que sont les systèmes de myriadisation ?

Le concept de myriadisation est venu de l'idée qu'un certain nombre de tâches pouvaient être effectuées par des utilisateurs d'Internet, en utilisant les atouts propres à celui-ci, c'est-à-dire pouvoir accéder à un grand nombre de personnes, de manière quasi-instantanée, partout dans le monde. La participation de ces internautes peut être bénévole ou rétribuée, suivant les tâches et les systèmes. Parmi les systèmes bénévoles, nous pouvons citer l'exemple fameux de Wikipedia et, parmi ceux avec rétribution, RentACoder (où l'on peut soumettre un projet de programmation à une communauté de programmeurs) ou LiveOps (qui est un centre d'appels virtuel, les opérateurs étant des internautes). A la suite de ces systèmes est apparu le concept de *Human computing*. Dans ce dernier cas, on ne fait plus appel à des compétences particulières d'internautes, mais on utilise deux propriétés très élémentaires : être un humain et avoir du temps libre. C'est l'application des grilles de calcul aux humains : chaque utilisateur, à la manière d'un processeur, effectue une tâche élémentaire en n'ayant accès qu'à la seule information nécessaire pour la mener à bien. Dans ce type de systèmes, seules des tâches très simples sont effectuées par les humains, soit parce qu'elles sont intrinsèquement simples (par exemple, mettre une étiquette sur une image), soit parce que la tâche est découpable en micro-tâches élémentaires. Ce sont les systèmes de myriadisation du travail parcellisé, qui sont le cœur de cet article. Dans ce concept, il y a souvent rétribution [2], mais celle-ci peut-être non monétaire, comme dans certains GWAP (*Games with a purpose*) (von Ahn, 2006; Chamberlain *et al.*, 2008). La création d'Amazon Mechanical Turk en 2005 s'inscrit dans cette dernière catégorie de systèmes de m.t.p. avec rémunération, qui a été suivie par un grand nombre d'autres systèmes (Biewald, 2010), ceux-ci n'ayant pas acquis la même importance, en particulier en raison du nombre de personnes inscrites. Comme souvent pour les nouveaux usages issus du Web, on ressent à la fois une fascination pour la potentialité des m.t.p. et une méfiance en face de ces pratiques qui ne semblent pas avoir de réelles considérations pour le droit du travail. L'apparition des systèmes de myriadisation pose de nombreux problèmes, éthiques et philosophiques, abordés par exemple dans (Zittrain, 2008), mais également légaux (Felstiner, 2010). Elle soulève d'importantes questions : qu'est-ce que le travail ? qu'est-ce qu'une rétribution juste ? un être humain est-il assimilable à un ordinateur ? Ces questions essentielles débordent largement à la fois le cadre d'un article de conférence et nos compétences. C'est pourquoi nous nous limiterons, autant que possible, aux problèmes précis que pose l'introduction de MTurk comme moyen de produire des ressources linguistiques, car nous jugeons cela à la fois urgent et crucial.

3 Amazon Mechanical Turk : légendes et réalité

Amazon Mechanical Turk (MTurk) permet, selon de nombreux auteurs dont le premier est Snow *et al.* (2008), de produire à peu de frais et rapidement des ressources linguistiques de qualité. Cette découverte est d'une telle importance pour la communauté qui manque toujours cruellement de moyens pour développer lesdites ressources, qu'elle a entraîné un important effet de mode. Ce phénomène est, nous allons le voir, ni totalement justifié, ni sans conséquences pour le développement futur de telles ressources. Par ailleurs, pour de nombreux chercheurs, les Turkers [3] utilisent MTurk comme un hobby, il n'est donc pas scandaleux de très mal les rémunérer. Nous allons voir ici que la situation est loin d'être aussi simplement idéale.

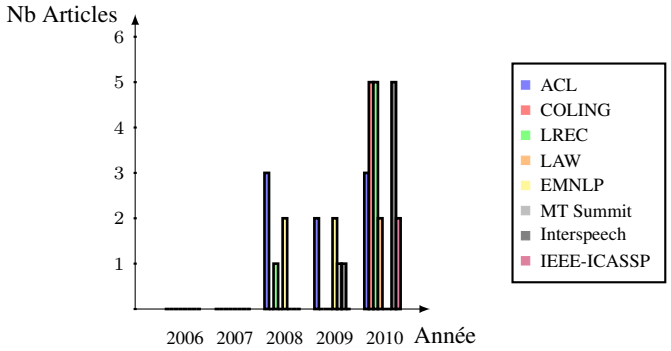
2. mais pas toujours, par exemple dans le système reCAPTCHA `http://www.google.com/recaptcha/learnmore`, où les CAPTCHAs proviennent de mots mal reconnus lors de la numérisation de Google books

3. Il est d'usage d'appeler les personnes effectuant des tâches au sein du « turc mécanique » des *Turkers*, et celles qui fournissent les tâches des *Requesters*.

3.1 Etat des lieux

Créé en 2005, le système de m.t.p. MTurk est aujourd'hui de plus en plus utilisé pour la création ou la validation de ressources linguistiques pour le TAL (Traitement Automatique des Langues), et la plupart des conférences internationales du domaine ont vu la présentation de projets de recherche utilisant MTurk.

Le graphique ci-dessous, repris de (Fort *et al.*, 2011), montre l'évolution rapide du phénomène. Il comptabilise le nombre de publications dans les principales conférences internationales décrivant des expériences utilisant MTurk.



Evolution de l'utilisation réelle de MTurk dans les publications TAL et parole

Afin de compléter cette étude, nous avons également réalisé une recherche similaire mais plus globale, cette fois dans l'anthologie de l'ACL. [4] Cette recherche, effectuée le 5 novembre 2010, a ramené 124 résultats, dont, après filtrage manuel, 86 papiers utilisant effectivement MTurk (Fort *et al.*, 2011). Ces résultats incluent un atelier spécialisé, fournissant 35 des 86 publications, le NAACL-HLT 2010 Workshop on Amazon Mechanical Turk, dont l'existence même est le signe de l'importance grandissante de MTurk dans le domaine. Mentionnons enfin qu'un certain nombre d'expériences relatées dans des articles ont utilisé MTurk sans le mentionner explicitement (Fort *et al.*, 2011). Côté francophone, nous ne trouvons aucun article utilisant MTurk dans les actes des précédents TALN, ni dans les numéros publiés à ce jour de la revue TAL.

3.2 MTurk est un hobby pour les Turkers ?

Afin de pouvoir efficacement juger de l'éthique et de la légalité de MTurk, il est fondamental de pouvoir qualifier l'activité que mène les Turkers lorsqu'ils effectuent des tâches dans MTurk. S'agit-il d'une activité bénévole, comme celle effectuée par les participants à Wikipedia ? clairement non, lorsque l'on regarde la page d'accueil où MTurk met directement l'accent sur l'argent gagné. Peut-elle être assimilée à un hobby, la rétribution étant alors assimilable à un bonus ne correspondant pas à un salaire, comme cela est suggéré dans quelques articles (Gao & Vogel, 2010; Novotney & Callison-Burch, 2010) ?

Un certain nombre d'études (Ross *et al.*, 2009, 2010; Ipeirotis, 2010b), fournissent, grâce à des questionnaires soumis aux Turkers *via* MTurk, des chiffres déclarés sur un certain nombre de facteurs socio-économiques (pays, âge, revenu, éducation...), sur la façon dont ils utilisent MTurk (nombre de tâches effectuées par semaine, revenu acquis, date d'entrée dans MTurk ...) [5] et dont ils qualifient leur activité. La motivation financière (déclarée) est minoritaire chez les Turkers américains (38%), mais majoritaire chez les Turkers indiens (69%). Si 60% des Turkers pensent que MTurk est un moyen utile de gagner de l'argent sur leur temps libre, ils ne sont que 30% à motiver leur participation par l'intérêt des tâches, et 20% (5% des travailleurs indiens) disent l'utiliser pour tuer le temps. Enfin, ils sont 20% (30% des Indiens) à dire que MTurk leur est nécessaire pour vivre, et à peu près le même pourcentage à dire que MTurk constitue leur principale source de revenus.

4. Association for Computational Linguistics, `http://www.aclweb.org/anthology/`

5. On pourra se reporter par exemple à (Adda & Mariani, 2010) pour un résumé de celles-ci. On y apprend par exemple (Ross *et al.*, 2010) que les Turkers en provenance d'Inde représentaient 5% à la fin de 2008, 36% fin 2009, plus de 50% en mai 2010 selon `http://blog.crowdflower.com/2010/05/amazon-mechanical-turk-survey/` et, selon (Biewald, 2010) sont responsables de plus de 60% de l'activité dans MTurk.

Un autre moyen de vérifier la nature de l'activité des Turkers est d'examiner la nature de la tâche. En effet, certaines tâches actuellement proposées sur MTurk correspondent à de nouveaux usages (par exemple, des expériences artistiques comme `http://www.thesheepmarket.com/`), mais d'autres étaient auparavant effectuées par des employés [6], et constituent donc un travail. Tel est le cas, par exemple, des activités de transcription ou de traduction, qui sont (pour ce qui concerne les ressources les plus significatives) produites par des employés d'entités comme le LDC ou ELDA. Pour les 20% des Turkers qui passent plus de 15h par semaine sur MTurk (Deneme, 2009; Adda & Mariani, 2010) et contribuent à hauteur de 80% des tâches, la durée d'activité est significative, et est assimilable à un travail.

Nous ne pouvons pas conclure de manière définitive sur la nature de l'activité de *tous* les Turkers, car la nature des tâches sur MTurk et la motivation des Turkers est composite. Cependant, nous pensons que pour les 20% des Turkers pour qui MTurk constitue la source principale de revenus, ainsi que pour les tâches assimilables à un travail (qui ont 8 chances sur 10 d'être effectuées par des Turkers travaillant plus de 15 heures par semaine), la nature de l'activité est assimilable à un travail.

3.3 MTurk permet de réduire les coûts ?

Dans la plupart des articles ayant utilisé MTurk, le faible coût de développement de la ressource est mis en avant. Il est vrai que MTurk permet de proposer des rétributions si faibles aux Turkers que le coût en est forcément réduit, par exemple 0.005$ pour transcrire un segment d'environ 5 secondes de parole téléphonique (Novotney & Callison-Burch, 2010). Il faut cependant nuancer ces chiffres. Tout d'abord, le coût effectif n'est pas toujours calculé avec rigueur. En effet, le temps de développement de l'interface et de mise en place des garde-fous est non nul (Callison-Burch & Dredze, 2010). De même, le coût de validation (Kaisser & Lowe, 2008) ou de développement (Xu & Klakow, 2010) post-MTurk permettant de compenser la mauvaise qualité des résultats (voir section 3.4) n'est généralement pas précisément évalué. Or, ces coûts supplémentaires ne sont jamais pris en compte dans le calcul final. De plus, certaines tâches peuvent se révéler plus coûteuses que prévues. Ainsi, si l'on ne trouve pas de Turkers pour faire la tâche, on peut être obligé d'augmenter la rémunération, comme Novotney & Callison-Burch (2010), qui, partant d'un coût très bas (5 dollars de l'heure transcrite), ont été obligés de le multiplier par 7 (37 dollars de l'heure) pour transcrire du coréen, par manque de Turkers qualifiés.

3.4 MTurk permet de produire une qualité équivalente ?

3.4.1 Limitations liées à la non expertise

Les Turkers étant des non experts, le Requester (fournisseur de tâches) doit découper les tâches complexes en tâches plus simples (HIT, Human Intelligence Task), afin de les rendre réalisables. Ce faisant, le chercheur est amené à faire des choix qui peuvent biaiser les résultats. Un exemple de ce type de biais est analysé dans (Cook & Stevenson, 2010), où les auteurs reconnaissent que le fait de ne proposer qu'une phrase par type d'évolution lexicale (amélioration ou péjoration) influence le résultat.

Plus grave encore que ces biais potentiels, certains chercheurs ont observé que, lorsque la complexité de la tâche augmente, la qualité produite sous MTurk est insuffisante. C'est notamment le cas dans (Bhardwaj *et al.*, 2010), qui démontre que, pour leur tâche de désambiguïsation lexicale, un petit nombre d'annotateurs bien formés produit de bien meilleurs résultats qu'un grand nombre (le nombre étant supposé contre-balancer la non expertise) de Turkers. De ce point de vue, leurs résultats contredisent ceux de Snow *et al.* (2008) dont la tâche était semblable mais beaucoup plus simple (nombre de sens proposé par mot de 3 au lieu de 9,5 dans (Bhardwaj *et al.*, 2010)). Cette même difficulté d'obtenir une qualité suffisante sur des tâches complexes apparaît dans (Gillick & Liu, 2010), qui démontre que l'évaluation par des non experts de systèmes de résumé automatique est « risquée », les Turkers n'étant pas capables d'obtenir des résultats comparables à ceux des experts. On retrouve ce problème de qualité dans de nombreux articles, dans lesquels les auteurs ont dû faire valider les résultats des Turkers par des spécialistes (des étudiants en thèse pour (Kaisser & Lowe, 2008)) ou leur faire subir un post-traitement assez lourd (Xu & Klakow, 2010). Enfin, la qualité du travail des annotateurs non experts varie considérablement (Tratz & Hovy, 2010).

Il existe également un effet « boule de neige » qui tend à surestimer la qualité signalée dans les articles : des chercheurs louent MTurk (Xu & Klakow, 2010), citant des recherches qui ont effectivement fait usage du système, mais

6. ce qui peut donc assimiler MTurk à une forme de délocalisation sur le web, pour faire baisser les coûts de production

qui n'auraient pas donné de résultats utilisables sans une intervention postérieure plus ou moins lourde (Kaisser & Lowe, 2008). Pire, lorsque MTurk sert à l'évaluation d'un système et lorsque son usage lui-même n'est pas évalué (Cook & Stevenson, 2010), on peut se demander quelle valeur attribuer à l'évaluation du système. On pourrait en conclure que MTurk ne devrait être utilisé que pour des tâches simples, or, outre le fait que son fonctionnement même induit d'importantes limitations (voir section suivante), il est intéressant de noter que dans certains cas simples, justement, des outils de TAL font d'ors et déjà mieux que les Turkers (Wais *et al.*, 2010).

3.4.2 Limitations liées au fonctionnement même de MTurk

Une première limitation est l'interface de MTurk. Tratz & Hovy (2010) notent ainsi que les limites de l'interface constituent « le premier et le plus important des défauts » de MTurk. Les auteurs regrettent par ailleurs l'impossibilité d'avoir la certitude que les Turkers participant à la tâche sont bien de langue maternelle anglaise. Cette impossibilité de connaître les capacités réelles des Turkers, notamment de connaître leur langue maternelle (bien que leurs adresses IP soient géolocalisables), est un problème bien réel. S'il est possible de mettre en place des tests préalables, qui, là encore, représentent un coût supplémentaire à prendre en compte, il est très facile de tricher (Callison-Burch & Dredze, 2010). Bien entendu, il est toujours possible de mettre en place des garde-fous (Callison-Burch & Dredze, 2010), mais, encore une fois, cela demande du temps et représente donc un coût supplémentaire que peu de Requesters sont prêts à investir. Ainsi, dans (Xu & Klakow, 2010), les auteurs ont identifié des spammeurs mais n'ont pas réussi à les éliminer. Pour certaines tâches, il peut s'avérer difficile de trouver des Turkers ayant les compétences nécessaires ((Gillick & Liu, 2010; Lambert *et al.*, 2010) en raison de la complexité de la tâche, (Gao & Vogel, 2010; Novotney & Callison-Burch, 2010) en raison de la langue à maîtriser).

Par ailleurs, il ne faut pas négliger l'impact du paiement à la tâche, qui induit comme comportement logique de placer le nombre de tâches réalisées au-dessus de la qualité de la réalisation, et ce, quelle que soit la rétribution. Kochhar *et al.* (2010) sont ainsi arrivés à la conclusion qu'il valait mieux payer à l'heure (avec, bien sûr, des procédures de vérification et de justification du temps passé).

4 Quelques réflexions sur le statut de MTurk

4.1 Quel est le statut de l'activité dans MTurk ?

En obscurcissant la relation entre Turkers et Requesters, et entre les Turkers eux-mêmes, MTurk empêche de fait la possibilité de s'organiser en syndicats, de protester contre d'éventuelles pratiques douteuses des Requesters ou d'ester en justice. Au-delà des problèmes de droit du travail, il faut parler des problèmes des taxes et cotisations sociales : Amazon considère (selon l'accord de licence de MTurk) que les Turkers sont assimilables à des travailleurs indépendants [7], et donc qu'il leur incombe de payer toutes les taxes et charges afférant à leur activité. Étant donné la hauteur des rémunérations prises individuellement, il est parfaitement hypocrite de penser que cela est possible. Il est donc fortement probable que les Turkers ne déclarent pas ces revenus et ne cotisent pas non plus à une quelconque caisse de retraite ou de sécurité sociale. Il en va bien entendu de même pour les fournisseurs de travail. Les états sont donc privés d'un revenu légitime.

Il faut souligner également que la nature de la relation entre les trois partenaires, Turker, Requester et MTurk, vague pour le droit américain, est encore plus douteuse en regard du droit français. En effet, selon la législation française du travail, en dehors du fait que le travail à la tâche est illégal, soit il s'agit de travail salarié, qui serait, en l'occurrence, non déclaré par l'employeur, donc illégal, soit il s'agit d'un rapport de prestation de service, dont le donneur d'ordre serait MTurk et le prestataire le Turker et, dans ce cas, le Turker doit être enregistré au registre du commerce.

4.2 Le modèle économique de MTurk est-il fondé ?

Comme souligné dans la partie 2, lorsque l'on aborde pour la première fois MTurk, on est sidéré devant les conditions financières imposées aux Turkers, qui amènent à des rémunérations horaires ridiculement basses (inférieures

7. Selon (Felstiner, 2010), cela n'est ni légal, ni défendable dans le droit américain, qui est le droit s'appliquant pour MTurk aux États-Unis.

à 2 dollars, soit 1,46 euros (Ross *et al.*, 2009; Ipeirotis, 2010b)). Ce coût fabuleusement bas correspond-il à une réalité économique saine (comme suggéré par Marge *et al.* (2010) et McGraw *et al.* (2010), qui considèrent que cette rétribution n'est qu'une sorte de bonus pour une tâche par ailleurs effectuée avec plaisir) ?

Nous l'avons vu dans la partie 3.2, l'assertion que les Turkers considèrent MTurk comme un hobby est fausse, au moins pour une partie significative d'entre eux. Dès lors, pourquoi, si cela constitue pour eux un travail, acceptent-ils un salaire horaire aussi bas ? la loi de l'offre et la demande n'est pas suffisante pour l'expliquer, tout d'abord parce que le nombre réel de Turkers n'est pas si important (Fort *et al.*, 2011), ensuite, parce qu'il est souvent difficile de faire exécuter des tâches de grande taille en un temps limité pour un coût standard (Ipeirotis, 2010a).

Un fait peut nous mettre sur la voie d'une explication crédible : beaucoup d'articles (par exemple (Marge *et al.*, 2010)) soulignent que la qualité n'est pas liée au coût associé à chaque tâche. Cela est dû en particulier à la présence de spammeurs (c'est-à-dire de Turkers qui répondent au hasard ou en utilisant un système automatique), attirés par les tâches bien rémunérées, et qui sont en grand nombre dans le système MTurk, le système de réputation mis en place par Amazon étant notoirement facile à mettre en défaut [8]. Cela conduit à une situation semblable au « marché des tacots », décrit par le prix Nobel Georges Akerlof (Akerlof, 1970) : l'acheteur d'une voiture d'occasion prend en compte dans le prix qu'il offre le risque que le vendeur lui « fourgue » un tacot. Les vendeurs propriétaires d'une bonne voiture ne peuvent donc obtenir un bon prix et quittent le système, ce qui accroît en retour la défiance de l'acheteur, car cela augmente le risque d'acheter un tacot. La présence des spammeurs, de par le laxisme du système mis en œuvre par Amazon, conduit à une stabilisation à un prix très bas, les bons travailleurs quittant donc le système (70% des Turkers utilisent MTurk depuis moins de 6 mois (Ross *et al.*, 2009)).

De plus, il est indéniable qu'un certain nombre de Turkers utilisent MTurk comme moyen de divertissement : ceux-ci sont attirés par les tâches intéressantes, quelle que soit leur rémunération. Sur ces tâches, ils sont en concurrence avec les Turkers-travailleurs (qui, naturellement, souhaitent également faire des tâches intéressantes), ce qui conduit également à faire baisser le taux horaire moyen « acceptable » [9].

Dernier facteur qui tend à faire accepter un taux horaire finalement inacceptable : le travail à la tâche. Un Turker, fait bien sûr une relation entre la difficulté de la tâche et la rétribution, mais n'a pas une idée claire du salaire horaire avant de commencer à travailler. De plus, le travail à la tâche induit un comportement que l'on peut voir également dans des jeux en ligne ou à chaque fois que l'on effectue une tâche contre une rétribution absolue : la personne a tendance à regarder grossir son compteur d'argent, ou de points, et à se fixer des objectifs absolus, déconnectés d'un quelconque taux horaire : « aujourd'hui, je reste dans le système, jusqu'à ce que j'ai gagné 5 dollars ». Ce qui n'est bien sûr pas le meilleur moyen d'optimiser le taux horaire. Si le travail à la tâche est interdit en France, c'est bien pour empêcher que des travailleurs gagnent moins que le salaire minimum.

Comme le souligne (Ipeirotis, 2010c), les défauts de la plateforme MTurk remettent en cause sa viabilité à moyen terme, si elle n'évolue pas fondamentalement, en particulier sur les problèmes de rémunération et de systèmes de réputation fiables pour les Turkers et les Requesters.

4.3 Quelle est la situation par rapport à la propriété intellectuelle sur MTurk ?

Au regard du droit européen, la problématique de la propriété intellectuelle pour des données telles que l'on peut en produire via MTurk se pose le plus souvent en termes de protection associée aux bases de données, au sens de la directive européenne du 11 mars 1996. Cette directive, qui concerne les ensembles d'informations de toutes natures, offre une double protection : (1) par le droit d'auteur concernant la structure de la base, conditionné au fait qu'il y ait là une *création originale*, et (2) par un droit spécifique couvrant le contenu, droit proche de celui du droit d'auteur mais conditionné à la valeur économique des données (et non à leur originalité), au sens où ces données doivent avoir été obtenues grâce à un *investissement substantiel du point de vue qualitatif ou quantitatif*. Cette deuxième protection est indépendante du caractère public ou non des données, l'objet de la protection étant la base dans son ensemble (c'est-à-dire l'assemblage des données).

Dans le cas de MTurk, les droits semblent devoir être la propriété du Requester, soit en tant qu'auteur (pour les HIT eux-mêmes et ce qu'ils pourraient contenir, sauf lorsque sont utilisées des données elles-mêmes soumises à des droits d'auteurs propres, comme, par exemple, si l'on fait transcrire ou traduire des contenus existants), soit en tant qu'organisateur de ce qui est une *œuvre collective* [10] (pour les productions des Turkers).

8. http ://behind-the-enemy-lines.blogspot.com/2010/10/be-top-mechanical-turk-worker-you-need.html

9. c'est-à-dire le taux horaire seuil, en dessous duquel un travailleur n'acceptera pas d'effectuer la tâche

10. L'œuvre collective est définie par l'article L. 113-2 alinéa 3 du Code de la propriété intellectuelle comme étant une œuvre *créée sur l'initiative d'une personne physique ou morale qui l'édite, la publie et la divulgue sous sa direction et son nom et dans laquelle la*

Cela dit, il n'est pas clair, avec MTurk relevant des États-Unis et des Requesters et des Turkers relevant souvent d'autres pays, qu'il y ait un droit applicable, la situation ne semblant pas envisagée dans les traités internationaux.

5 Alternatives existantes ou proposées

Comme indiqué ci-dessus, les objectifs principaux des développeurs de ressources linguistiques faisant appel à MTurk sont l'obtention de résultats de bonne qualité, à un faible coût et dans un délai très bref. Mais ces objectifs ne sont pas nécessairement faciles à atteindre avec MTurk, alors que des approches alternatives existent. Tout d'abord, bien qu'une comparaison systématique entre MTurk et des algorithmes état-de-l'art impliquant un échantillon varié de tâches liées au TAL reste à faire, il semble que certains auteurs aboutissent à la conclusion que les annotateurs automatiques déjà disponibles pour certaines tâches font aussi bien voire mieux que les Turkers (Wais *et al.*, 2010) : les outils automatiques peuvent faire mieux que les non experts. La réutilisation intelligente de ressources existantes peut également être une alternative simple et peu coûteuse à MTurk. Enfin, MTurk n'est qu'une des nombreuses possibilités de m.t.p.

5.1 Approches non-supervisées et semi-supervisées pour le développement de ressources linguistiques à faible coût

La communauté du TAL s'intéresse depuis longtemps à des approches dites *non-supervisées* d'apprentissage automatique, pour un large éventail de tâches parfois complexes. De la segmentation en mots à l'analyse syntaxique (Hänig, 2010) en passant par l'annotation morphosyntaxique (Goldwater & Griffiths, 2007), le développement de ressources lexicales (y compris de niveau sémantique ou pragmatique, cf. (Pak & Paroubek, 2010)) ou la catégorisation de documents, nombreuses sont les tâches pour lesquelles des techniques existent qui ne nécessitent aucune ressource préalable. Bien que les résultats obtenus soient souvent inférieurs aux résultats des approches supervisées (utilisant un corpus d'apprentissage) ou symboliques avancées (utilisant des ressources symboliques également coûteuses à développer), il n'est pas clair qu'ils soient inférieurs à ce que l'on peut attendre de MTurk. C'est notamment le cas pour des tâches complexes comme l'analyse syntaxique.

Pour améliorer à faible coût la qualité des outils statistiques ainsi développés et/ou pour les faire correspondre à des modèles préexistants (par exemple, à un inventaire préétabli de catégories dans le cadre de l'annotation morphosyntaxique), il n'est pas forcément nécessaire de recourir à des techniques totalement supervisées. Une utilisation optimale d'un ensemble limité d'informations (annotations, ressources externes) peut donner de bons résultats : c'est le paradigme de l'*apprentissage semi-supervisé* (Abney, 2007). Dans le cas du développement de ressources linguistiques, on peut identifier deux types (non mutuellement exclusifs) de semi-supervision.

La première idée que l'on peut avoir est d'entraîner des modèles sur les quelques données annotées, puis d'annoter automatiquement les autres données : on peut alors choisir parmi les données annotées automatiquement celles pour lesquelles le modèle a un niveau de confiance optimal, et les considérer comme de nouvelles données annotées pour l'apprentissage d'un nouveau modèle, et ainsi de suite : c'est le *self-training*, utilisé en TAL depuis longtemps (Yarowsky, 1995). Cette idée peut être généralisée en utilisant deux modèles les plus différents possible [11], et à compléter les données d'apprentissage de l'un par les annotations automatiques les plus sûres produites par l'autre. C'est le *co-training* (Blum & Mitchell, 1998), qui cherche à éliminer au maximum les biais spécifiques à chaque modèle par la confrontation à un autre. À ce stade, on reste dans une situation où une annotation manuelle peu coûteuse sert de graine pour la construction successive, mais automatisée, de modèles qui vont en s'améliorant, jusqu'à obtenir des performances satisfaisantes. Si l'on accepte de continuer à annoter des données manuellement au cours des étapes de construction successive de modèles, on peut faire en sorte que soient choisies et présentées aux annotateurs les données telles que disposer d'une annotation de référence pour elles soit de nature à améliorer au mieux la qualité des outils. C'est l'idée au cœur de l'*active learning* (Cohn *et al.*, 1995).

La deuxième idée, combinable avec la première, consiste à utiliser au mieux des données annotées d'une façon moins complète que l'annotation visée. Par exemple, pour l'annotation morphosyntaxique, on peut disposer d'un lexique externe mais pas d'un corpus d'apprentissage : projeter le lexique sur le corpus correspond alors

contribution personnelle des divers auteurs participant à son élaboration se fond dans l'ensemble en vue duquel elle est conçue, sans qu'il soit possible d'attribuer à chacun d'eux un droit distinct sur l'ensemble réalisé. L'article 113-5 stipule alors que *L'œuvre collective est, sauf preuve contraire, la propriété de la personne physique ou morale sous le nom de laquelle elle est divulguée. [...]* .

11. Ce qui peut être difficile à construire (Ng & Cardie, 2003).

à une annotation ambiguë, qu'il faut désambiguïser (Smith & Eisner, 2005). Pour le développement de lexiques morphologiques, disposer d'une description formalisée de la morphologie de la langue permet l'utilisation de techniques efficaces de suggestion d'entrées lexicales (Sagot, 2005). De même, on peut chercher à exploiter un corpus partiellement parenthésé pour guider des modèles d'analyse syntaxique complets (Watson *et al.*, 2007).

5.2 Réutilisation de ressources existantes

Moins coûteux encore, la construction de ressources linguistiques peut se faire en réutilisant des ressources existantes. Considérons par exemple la tâche de détection d'entités nommées. Nothman *et al.* (2008) montrent qu'il est possible de transformer Wikipedia en une ressource annotée en entités nommées de large couverture et de très bonne qualité. De tels corpus ont pourtant été construits au moyen de MTurk, notamment sur des corpus non-standard, en particulier médicaux (Yetisgen-Yildiz *et al.*, 2010), twitter (Finin *et al.*, 2010), e-mails (Lawson *et al.*, 2010). Naturellement, ces corpus sont très différents de ce que l'on peut obtenir au moyen de Wikipedia. Mais la taille des données extraites, ainsi que les caractéristiques de Wikipedia en tant que corpus, font que les détecteurs d'entités nommées entraînés sur un corpus construit à partir de Wikipedia tendent à avoir de très bons résultats lorsqu'ils sont utilisés sur d'autres types de corpus (Balasuriya *et al.*, 2009).

Il ne s'agit là que d'un exemple, mais nombreuses sont les ressources susceptible de fournir des données de toutes natures : Wikipedia [12] et autres projets wiki, notamment wiktionary, corpus (annotés ou non, oraux ou textuels) et ressources lexicales (phonétiques, morphologiques, syntaxiques, sémantiques), pour peu qu'elles soient disponibles pour la communauté. Il s'agit ici d'un autre débat, sur lequel nous n'insisterons donc pas plus avant.

5.3 Développement collaboratif ou myriadisé de ressources linguistiques au-delà de MTurk

Toutes les méthodes alternatives décrites jusqu'à présent ont prouvé leur utilité et leur efficacité, mais elles requièrent des compétences expertes, ne serait-ce que pour concevoir et développer les outils automatiques, mais également pour effectuer, si besoin est, les tâches d'annotation optimisées. Il existe des méthodes de développement de ressources linguistiques qui ne font pas nécessairement appel à des experts, sans être pour autant touchées par tous les problèmes décrits pour MTurk. Il s'agit en particulier des approches collaboratives, des approches ludiques mais également de certaines plateformes de m.t.p., qui se sont données les moyens d'en éviter les écueils.

Les approches collaboratives de développement de ressources lexicales reposent sur la stratégie mise en place par le projet Wikipedia, les autres projets de la constellation Wikimedia, et d'autres types de wiki comme les *Semantic Wiki* (Freebase, OntoWiki…). Différents participants volontaires, experts ou non, enrichissent progressivement une même ressource, soit sous forme d'annotations soit sous forme de bases de données (lexicales, ontologiques…). C'est une première étape vers la m.t.p. : ici, le travail n'est pas parcellisé, et il n'est que faiblement myriadisé. Les annotations des uns sont « contrôlées » par les autres, et des divergences de vues entre différents participants se manifestent le plus souvent par des discussions, conduisant éventuellement à ce que l'administrateur tranche et décide. C'est ainsi qu'un très haut niveau de qualité peut être finalement atteint. Une des premières plateformes wiki dédiée au développement d'une ressource TAL est l'outil Serengeti, développé à l'Université de Bielefeld (Stürenberg *et al.*, 2007), à des fins d'annotation sémantique des textes. Cet outil est utilisé actuellement dans le cadre du projet AnaWiki (`http://www.anawiki.org`).

Toutefois, ces approches restent plus adaptées pour le développement de ressources de taille raisonnable avec une bonne qualité (*gold standard*). Elles sont moins indiquées pour le développement rapide de ressources à grande échelle. Une autre stratégie, qui repose également sur le Web, est d'attirer de grand nombres de non experts au moyen de jeux en ligne dit *ayant un but* (en anglais *games with a purpose*, ou *GWAP*). Cette idée, initiée par (von Ahn, 2006; von Ahn & Dabbish, 2008) avec le jeu en ligne ESP (`http://www.espgame.org/`) consiste à faire étiqueter des images par des joueurs qui rentrent en compétition : ceux-ci reçoivent des crédits lorsque leurs réponses coïncident avec celles d'autres joueurs [13]. ESP a connu un succès important en mobilisant $13,500$ utilisateurs, créant $1,3$ million d'étiquettes dans les premiers mois suivant son apparition sur la Toile. Cette idée a été par la suite déclinée pour diverses types de tâches, y compris en TAL. Des exemples en sont le jeu *JeuxDeMots* (Lafourcade & Joubert, 2008, `http://www.lirmm.fr/jeuxdemots`), qui vise à collecter des relations entre

12. Il faut notamment citer le projet DBpedia (`http://dbpedia.org`), qui cherche à extraire des informations ontologiques structurées à partir de Wikipedia, constituant ainsi une ressource aux potentialités multiples et déjà largement utilisée.

13. Différentes procédures sont prévues pour exclure les utilisateurs malveillants, notamment le contrôle des adresses IP, la vérification aléatoires des étiquetages pour des réponses connues, etc. (von Ahn & Dabbish, 2008)

mots, et son *alter ego* PtiClic (ibid., http://www.lirmm.fr/pticlic), qui vise à typer explicitement ces relations. Le jeu *PhraseDetective* (Chamberlain *et al.*, 2008, http://www.phrasedetectives.org), quant à lui, a pour objectif l'annotation de liens anaphoriques, tâche pourtant réputée complexe. L'idée est alors que l'on peut aussi utiliser le jeu pour former les utilisateurs à la tâche. *Phrase Detective* comprend ainsi une phase d'entraînement où l'on apprend la tâche au nouveau joueur, par le biais de tests de plus en plus durs basés sur un petit ensemble de données venant de corpus existants (annotés par des experts).

Toutefois, la frontière entre jeu de type GWAP et m.t.p. à la MTurk n'est pas nette. On ne peut pas distinguer facilement les GWAP, qui seraient plus ludiques, et MTurk, qui serait *stricto sensu* du travail : même contribuer à Wikipedia est un travail, certes bénévole. On ne peut pas non plus distinguer les GWAP de MTurk en tant qu'ils ne donneraient lieu à aucune récompense tangible : certains jeux en ligne sont des GWAP mais proposent des rémunérations non-monétaires (ainsi, PhraseDetective permet de gagner des bons à dépenser sur le site d'achat en ligne Amazon). Enfin, on ne peut pas distinguer la m.t.p. à la MTurk par le caractère « éthique » des premiers. Il existe en effet des alternatives à MTurk pour développer des ressources linguistiques dans le paradigme de la m.t.p., tout en évitant les écueils évoqués tout au long de cet article.

Pour le recueil de données langagières, en particulier pour les langues peu dotées, des alternatives à MTurk semblent être plus appropriées. Ainsi, l'utilisation d'applications sur des téléphones portables de nouvelle génération est un moyen plus efficace de pouvoir accéder à toute une population. Hughes *et al.* (2010) ont ainsi embauché des locuteurs locaux et leur ont prêté des téléphones sur lesquels tournait une application dédiée. Les auteurs ont ainsi recueilli 3 000 heures en 17 langues. Un exemple de plateforme éthique de m.t.p. est Samasource, une ONG qui utilise ce type de méthode pour faire effectuer des tâches [14] à des personnes réellement nécessiteuses [15], formées pour cette tâche. Il s'agit là d'une alternative éthique à l'utilisation de MTurk qui permet également de tirer parti des avantages de la m.t.p.

5.4 Optimiser le coût de l'annotation manuelle : pré-annotation et interfaces dédiées

Indépendamment de la façon dont on s'en sert, l'annotation manuelle par des experts peut être considérablement accélérée voire améliorée au moyen d'outils d'annotation automatique utilisés comme pré-annotateurs. Par exemple, Fort & Sagot (2010) ont démontré que, dans le cas de l'étiquetage morphosyntaxique, une préannotation, même de piètre qualité et donc développable à faible coût, permet d'améliorer très largement le temps et la qualité des annotations manuelles. Ainsi, les auteurs ont montré que 50 phrases annotées à la main sans pré-annotation, ce qui prend environ 40 minutes [16], permettent de construire un préannotateur tel que la vitesse de l'annotation manuelle par un expert est quasiment identique à ce que l'on obtient avec un préannotateur de niveau état-de-l'art, c'est-à-dire que l'on peut construire un corpus complet de taille standard (10 000 phrases) en environ 6 000 minutes (100 heures). Des annotateurs experts et coûteux, pour peu que leur travail soit préparé puis utilisé de façon optimale, permettent donc le développement de ressources de très bonne qualité à un coût qui reste limité. À l'inverse, sur cette tâche d'apparence simple, des Turkers seraient bien en peine de suivre correctement un guide d'annotation détaillé, nécessairement complexe s'il est linguistiquement sérieux.

Par ailleurs, les remarques de Tratz & Hovy (2010) mentionnées ci-dessus concernant les limitations des interfaces déployables dans MTurk s'appliquent de manière générale. L'expérience acquise, par exemple, dans le développement de corpus annotés syntaxiquement ou sémantiquement montre que la rapidité et la qualité l'annotation, de quelque nature qu'elle soit, est fortement influencée par l'interface d'annotation elle-même (cf. par exemple (Erk *et al.*, 2003)). Il y a donc là aussi matière à accélérer et améliorer toute étape d'annotation manuelle, au point qu'une interface adaptée à une tâche donnée pourrait permettre de réduire les coûts dans des proportions comparables à celles obtenues par l'utilisation de MTurk, sans en présenter les inconvénients.

6 Conclusion et perspectives

MTurk illustre la complexité et la difficulté d'appréhender les relations (commerciales, de travail et autres) dans les nouveaux modes d'activités sur Internet. Les chercheurs qui ont utilisé MTurk l'ont fait souvent de bonne foi,

14. Comme traduire des SMS en créole, lors du tremblement de terre à Haïti afin de permettre aux secours internationaux d'aller à leur secours. http://www.samasource.org/haiti/.

15. Et rémunérées équitablement selon des barèmes dépendant du pays.

16. Les estimations proposées dans ce paragraphe, très grossières, reposent sur celles de (Fort & Sagot, 2010).

pour produire plus de données et les redistribuer à la communauté. Pour ceux qui ont eu des doutes sur de possibles problèmes d'éthique et de droit du travail, une recherche superficielle les a convaincu que MTurk est une sorte d'avatar de Wikipedia, et que les Turkers sont motivés surtout par le plaisir d'effectuer des tâches amusantes.

Nous pensons avoir montré que MTurk n'est pas une panacée et que d'autres solutions existent aujourd'hui pour réduire les coûts de construction de ressources linguistiques de qualité, tout en respectant ceux qui travaillent sur ces ressources et en tirant un meilleur parti de leurs compétences. Car derrière le débat autour de MTurk se trouve finalement la question de la considérationdue aux annotateurs, aux traducteurs, aux spécialistes de la transcription.

Nous aimerions, en conclusion, aller au-delà des faits actuels et mettre l'accent sur les conséquences à plus ou moins long terme de cette « mode ». En effet, sous la pression de ce type de systèmes à bas coût, les agences de moyens pourraient bientôt être plus réticentes à financer des projets de développement de ressources linguistiques à des coûts « normaux » (ou plutôt réalistes). Le coût à la MTurk deviendrait alors une norme de fait et nous n'aurions plus le choix de nos méthodes de développement.

Nous avons vu, dans la partie 5.3, qu'un système de m.t.p. peut permettre de faire produire des tâches rémunérées en préservant l'éthique, cela peut même être une chance pour des personnes qui ne peuvent se trouver sur le marché du travail, de par leur isolement, leur handicap, etc ; mais cela nécessite un encadrement légal strict afin de s'assurer que ce système n'est pas une remise en cause des droits des travailleurs. C'est pourquoi nous proposons la création d'un label de qualité et d'éthique, qui pourrait être décerné aux ressources par les associations savantes concernées, l'ATALA [17] pour le TAL et l'AFCP [18] pour la parole. Les questions d'éthiques sont dès à présent un critère de sélection pour les projets européens, ce label permettrait de préciser le statut des ressources comme critère de sélection pour l'ensemble des agences de moyens, tout en valorisant les bonnes pratiques de développement.

Remerciements

Ce travail a été réalisé en partie dans le cadre du programme Quaero, financé par OSEO, agence nationale de valorisation de la recherche, et dans celui du projet ANR EDyLex (ANR-09-CORD-008).

Références

ABNEY S. (2007). *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC, 1ère edition.

ADDA G. & MARIANI J. (2010). Language resources and amazon mechanical turk : legal, ethical and other issues. In *LISLR2010, "Legal Issues for Sharing Language Resources workshop", LREC2010*.

AKERLOF G. A. (1970). The market for 'lemons' : Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, **84**(3), 488–500.

BALASURIYA D., RINGLAND N., NOTHMAN J., MURPHY T. & CURRAN J. R. (2009). Named entity recognition in wikipedia. In *People's Web '09 : Proceedings of the 2009 Workshop on The People's Web Meets NLP*, p. 10–18, Morristown, NJ, USA : Association for Computational Linguistics.

BHARDWAJ V., PASSONNEAU R., SALLEB-AOUISSI A. & IDE N. (2010). Anveshan : A tool for analysis of multiple annotators' labeling behavior. In *Proceedings of The fourth linguistic annotation workshop (LAW IV)*, Uppsala, Suède.

BIEWALD L. (2010). Better crowdsourcing through automated methods for quality control. *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*.

BLUM A. & MITCHELL T. (1998). Combining labeled and unlabeled data with co-training. In *COLT : Proceedings of the Workshop on Computational Learning Theory* : Morgan Kaufmann Publishers.

CALLISON-BURCH C. & DREDZE M. (2010). Creating speech and language data with amazon's mechanical turk. In *CSLDAMT '10 : Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Morristown, NJ, USA : Association for Computational Linguistics.

CHAMBERLAIN J., POESIO M. & KRUSCHWITZ U. (2008). Phrase Detectives : a Web-based Collaborative Annotation Game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz.

17. http://www.atala.org/
18. http://www.afcp-parole.org/

COHN D. A., GHAHRAMANI Z. & JORDAN M. I. (1995). Active learning with statistical models. In G. TESAURO, D. TOURETZKY & T. LEEN, Eds., *Advances in Neural Information Processing Systems*, volume 7, p. 705–712 : The MIT Press.

COOK P. & STEVENSON S. (2010). Automatically identifying changes in the semantic orientation of words. In N. C. C. CHAIR), K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* : European Language Resources Association (ELRA).

DENEME (2009). How many turkers are there ? http ://groups.csail.mit.edu/uid/deneme/ ?p=502.

ERK K., KOWALSKI A. & PADO S. (2003). The salsa annotation tool. In D. DUCHIER & G.-J. M. KRUIJFF, Eds., *Proceedings of the Workshop on Prospects and Advances in the Syntax/Semantics Interface*, Nancy, France.

FELSTINER A. (2010). Working the Crowd : Employment and Labor Law in the Crowdsourcing Industry. *SSRN eLibrary*.

FININ T., MURNANE W., KARANDIKAR A., KELLER N., MARTINEAU J. & DREDZE M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 80–88, Stroudsburg, PA, USA : Association for Computational Linguistics.

FORT K., ADDA G. & COHEN K. B. (2011). Amazon mechanical turk : Gold mine or coal mine ? *Computational Linguistics (editorial)*, **37**(2).

FORT K. & SAGOT B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. In *Proc. of the Fourth ACL Linguistic Annotation Workshop*, Uppsala, Suède.

GAO Q. & VOGEL S. (2010). Consensus versus expertise : A case study of word alignment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 30–34, Stroudsburg, PA, USA : Association for Computational Linguistics.

GILLICK D. & LIU Y. (2010). Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 148–151, Stroudsburg, PA, USA : Association for Computational Linguistics.

GOLDWATER S. & GRIFFITHS T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL*, Prague, République tchèque.

HÄNIG C. (2010). Improvements in unsupervised co-occurrence based parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.

HUGHES T., NAKAJIMA K., HA L., VASU A., MORENO P. & LEBEAU M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. In *Proceedings of Interspeech*, p. 1914–1917.

IPEIROTIS P. (2010a). Analyzing the amazon mechanical turk marketplace. CeDER Working Papers, http ://hdl.handle.net/2451/29801. CeDER-10-04.

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

IPEIROTIS P. (2010c). A plea to amazon : Fix mechanical turk ! http ://behind-the-enemy-lines.blogspot.com/2010/10/plea-to-amazon-fix-mechanical-turk.html.

KAISSER M. & LOWE J. B. (2008). Creating a research collection of question answer sentence pairs with amazon's mechanical turk. In *Proceedings of the International Language Resources and Evaluation (LREC-2008)*.

KOCHHAR S., MAZZOCCHI S. & PARITOSH P. (2010). The anatomy of a large-scale human computation engine. In *Proceedings of Human Computation Workshop at the 16th ACM SIKDD Conference on Knowledge Discovery and Data Mining, KDD 2010*, Washington D.C.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, p. 657–666.

LAMBERT B., SINGH R. & RAJ B. (2010). Creating a linguistic plausibility dataset with non-expert annotators. In *Proceedings of Interspeech*, p. 1906–1909.

LAWSON N., EUSTICE K., PERKOWITZ M. & YETISGEN-YILDIZ M. (2010). Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 71–79, Stroudsburg, PA, USA : Association for Computational Linguistics.

MARGE M., BANERJEE S. & RUDNICKY A. I. (2010). Using the amazon mechanical turk for transcription of spoken language. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, p. 5270–5273, Dallas, TX.

MCGRAW I., YING LEE C., HETHERINGTON L., SENEFF S. & GLASS J. (2010). Collecting voices from the cloud. In *Proceedings of the International Language Resources and Evaluation (LREC-2010)*, p. 1576–1583.

NG V. & CARDIE C. (2003). Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of EMNLP 2003*.

NOTHMAN J., CURRAN J. R. & MURPHY T. (2008). Transforming Wikipedia into Named Entity Training Data. In *Proceedings of the Australian Language Technology Workshop*, p. 124–132.

NOVOTNEY S. & CALLISON-BURCH C. (2010). Cheap, fast and good enough : automatic speech recognition with non-expert transcription. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, p. 207–215, Stroudsburg, PA, USA : Association for Computational Linguistics.

PAK A. & PAROUBEK P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, La Valette, Malte : European Language Resources Association (ELRA).

ROSS J., IRANI L., SILBERMAN M. S., ZALDIVAR A. & TOMLINSON B. (2010). Who are the crowdworkers ? : shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, p. 2863–2872, New York, NY, USA : ACM.

ROSS J., ZALDIVAR A., IRANI L. & 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. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 ((c) Springer-Verlag), Proceedings of TSD'05*, p. 156–163, Karlovy Vary, République tchèque.

SMITH N. & EISNER J. (2005). Contrastive estimation : Training log-linear models on unlabeled data. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 354–362, nn Arbor, Michigan, USA.

SNOW R., O'CONNOR B., JURAFSKY D. & NG. A. Y. (2008). Cheap and fast - but is it good ? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, p. 254–263.

STÜRENBERG M., GOECKE D., DIE-WALD N., CRAMER I. & MEHLER A. (2007). Web-based annotation of anaphoric relations and lexical chains. In *ACL Workshop on Linguistic Annotation Workshop (LAW)*, Prague, République tchèque.

TRATZ S. & HOVY E. (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 678–687, Uppsala, Suède : Association for Computational Linguistics.

VON AHN L. (2006). Games with a purpose. *IEEE Computer Magazine*, p. 96–98.

VON AHN L. & DABBISH L. (2008). General techniques for designing games with a purpose. *Communications of the ACM*, p. 58–67.

WAIS P., LINGAMNENI S., COOK D., FENNELL J., GOLDENBERG B., LUBAROV D., MARIN D. & SIMONS H. (2010). Towards building a high-qualityworkforce with mechanical turk. In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*.

WATSON R., BRISCOE T. & CARROLL J. (2007). Semi-supervised training of a statistical parser from unlabeled partially-bracketed data. In *Proceedings of the 10th International Conference on Parsing Technologies*, IWPT '07, p. 23–32, Stroudsburg, PA, USA : Association for Computational Linguistics.

XU F. & KLAKOW D. (2010). Paragraph acquisition and selection for list question using amazon's mechanical turk. In *Proceedings of the International Language Resources and Evaluation (LREC-2010)*, p. 2340–2345, La Valette, Malte.

YAROWSKY D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, p. 189–196, Cambridge, MA.

YETISGEN-YILDIZ M., SOLTI I., XIA F. & HALGRIM S. R. (2010). Preliminary experience with amazon's mechanical turk for annotating medical named entities. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 180–183, Stroudsburg, PA, USA : Association for Computational Linguistics.

ZITTRAIN J. (2008). Ubiquitous human computing. *Phil. Trans. R. Soc. A 28*, **366**(1881), 3813–3821.