Sciences participatives et TALN: jusqu'où ? comment ? pourquoi ?

Jean-Yves Antoine[1], Anaïs Lefeuvre-Halftermeyer[2, 1]
(1) Université François Rabelais Tours, LI, 41000 Blois
(2) Université d'Orléans, LIFO 45000 Orléans
Jean-Yves.Antoine@univ-tours.fr, Anais.Halftermeyer@univ-orleans.fr

RESUME

Cet article s'interroge sur les modalités de participation citoyenne aux recherches en TALN, à la lumière des projets actuels en sciences citoyennes mais aussi d'études menées sur le sujet en histoire des sciences. Il vise à montrer comment une science participative est déjà en marche en TALN, à interroger ses modalités et également à en circonscrire les limites.

ABSTRACT

Citizen science and NLP : how far ? how ? why ?
This paper investigates the modalities of achievement of citizen science in NLP, by considering existing participative projects but also historical studies on the relationships between science and opinion. It questions the benefits but also the limitations of the involvement of citizens in NLP researches and advocates for the experimentation of such participation projects.

MOTS-CLES : Ethique, science participative, crowdsourcing, évaluation.
KEYWORDS: Ethics, Citizen science, crowdsourcing, evaluation.

1 Introduction

Les recherches menées en TALN s'appuient sur l'utilisation de ressources linguistiques comme matière première, et ce selon une certaine variété de méthodes. La constitution de ces ressources peut répondre à différentes finalités : soit la récupération de données linguistiques sources réelles (corpus brut), soit l'enrichissement de ces dernières par une annotation fine portant sur certains phénomènes spécifiques.

La constitution de ces ressources linguistiques implique de fait le citoyen (dit « naïf »[1] de la discipline, par opposition au chercheur en TALN), que ce soit pour produire des données premières, enrichir ces données ou bien pour faire émerger de nouveaux besoins. Les travaux faisant intervenir ces citoyens-locuteurs sont nombreux aujourd'hui, dans la constitution de corpus,

[1] Pour une discussion sur la notion d'annotateur expert, non-expert ou naïf, on pourra se référer à l'état de l'art donné par (Fort 2017) dans le paragraphe 2.2.3 (pages 87 à 93). Par ailleurs, nous utilisons le terme « citoyen » pour désigner l'ensemble de la population qui n'est pas professionnellement dans le domaine de la recherche (le chercheur-citoyen sera nommé abusivement « chercheur »). Ce choix est apparu naturel dans l'idée d'une science citoyenne, et à défaut de terme dédié.

(annotés ou non), de lexiques ou dictionnaires (cf. *Wiktionary*[2] pour un exemple multilingue), voire de grammaires. Il semble important de souligner dès à présent que ces « naïfs » participent par ce biais déjà à l'évolution du TALN lui-même. Ces interventions répondent à deux modalités de participation très distinctes :

- Soit les locuteurs produisent des ressources volontairement et/ou à dessein, et sont donc informés de la finalité de leur participation : ce type de participation se rencontre avec les jeux sérieux[3] tels que *ZombiLingo* (Fort *et al.* 2014, Guillaume et al. 2016) et *JeuxdeMots* (Lafourcade et Joubert 2008), pour les plus connus en France, les plateformes de contribution telles que *Sms4Science*[4] dont une déclinaison française est *Sud4Science* (Pankhurst *et al.* 2013), ou encore les travaux sur l'annotation de phrase en polarité émotionnelle (Antoine *et al.* 2014, par exemple).

- Soit le chercheur en TALN rassemble des productions diffusées à l'origine dans d'autres buts que de supporter la recherche (corpus de journaux, de blog etc.), et donc collectées sans information du citoyen sur ce nouvel usage.

Ces deux types d'intervention citoyenne diffèrent de deux points de vue :

- celui du *consentement éclairé* de l'utilisation de données par leur producteur, qui n'est assuré que dans le premier cas. Ce consentement s'accompagne d'ailleurs d'une démarche volontaire et active du citoyen.

- Celui de la *nature des informations collectées* : alors que dans le cas d'un enrichissement par annotation, le producteur crée consciemment une vraie connaissance originale, dans le cas de la collecte de corpus bruts, celui-ci est un simple émetteur de signaux, sa situation se rapprochant dès lors de celle des sujets d'expérimentations en médecine, en psychologie etc. (consentement éclairé en moins).

Comment situer ces interventions du grand public en TALN dans la question plus générale des sciences participatives ? Débat sur les cultures transgéniques, commissions de consensus sur les nanotechnologies, controverses sur les radio-émissions, condamnation d'experts sismologues, les rapports entre science et grand public sont de plus en plus empreints d'une grande ambivalence. La science est en effet vue comme le facteur de progrès dans notre société technique, ou au contraire comme une menace, ou une autorité trop liée aux pouvoirs dominants. Dans ces débats, le chercheur voit sa légitimité contestée. Les sciences participatives proposent un recours à la dégradation des rapports de confiance entre science et opinion publique, et participent à la mise en œuvre d'une vraie démocratie technique (Callon *et al.* 2001). Quoiqu'encore modestes en France, les initiatives en faveur d'une science citoyenne se multiplient. En témoignent par exemple certains appels à programmes de recherche associant laboratoires et associations de citoyens (PICRIS[5] en Ile-de-France, Chercheurs Citoyens dans l'ex-région Nord-Pas-de-Calais[6], le groupe sciences

---

2 https://www.wiktionary.org/

3 Dans le cas des jeux sérieux ou jeux avec un but – GWAP : Game With A Purpose (Fort, 2017), la motivation principale reste l'amusement, et le joueur peut simplement s'informer sur la finalité de sa participation s'il le souhaite, en suivant un lien. Cette finalité reste toutefois explicite : *Zombilingo* annonce ainsi « joue pour aider les scientifiques » sur sa page d'accueil.

4 Voir http://www.sms4science.org.

5 Voir www.iledefrance.fr/aides-regionales-appels-projets/partenariats-institutions-citoyens-recherche-innovation-picri.

6 Voir www.nordpasdecalais.fr/jcms/c_20218/guide-des-aides/programme-chercheurs-citoyens-2016.

participatives de l'alliance Athena[7]). Ce réinvestissement du citoyen dans les questions scientifiques concerne toutefois avant tout les sciences qui ont une visibilité politique forte, en particulier celles relevant de la santé ou de l'environnement. À un degré moindre, notre société du numérique voit émerger des interrogations citoyennes sur les technologies influant sur la structuration de la société.

Le TALN est pour l'heure peu concerné par ce mouvement, alors que ses applications commencent à envahir des champs importants de la vie quotidienne. L'application de la fouille de texte au monitoring des comportements sur les réseaux sociaux est un exemple, parmi d'autres, des questions éthiques que posent désormais le TALN (Lefeuvre *et al.* 2015). Le développement prévisible des technologies langagières devrait appeler une attente citoyenne de plus en plus marquée à laquelle les sciences participatives pourraient répondre.

Ces échanges citoyens autour des technologies langagières peuvent bénéficier d'un facteur favorisant : tout locuteur est porteur d'une expertise praticienne de sa langue maternelle, objet socio-linguistique à laquelle il contribue au quotidien. Cette expertise s'observe dans nos jeux de mots humoristiques ou nos pratiques diaphasiques (Gadet 2006). Sans être complètement conscientisée ou théorisée, elle relève en partie de notre apprentissage scolaire au cours duquel on est guidé vers une autoévaluation et une maîtrise de cette compétence, contrairement à d'autres aptitudes moins conscientisées telle que la marche par exemple. Le TALN constitue de fait un lieu prédisposé pour évaluer les possibilités du public à s'approprier les enjeux de la recherche scientifique. En éclairant la compréhension des questions scientifiques liées au progrès technique, la formation comme la vulgarisation sont de puissants outils pour atteindre cet objectif démocratique essentiel, même si, dès lors que l'on se place dans le cadre de la transmission du savoir, il persiste toujours un enjeu de pouvoir. Il est intéressant de se demander dans quelle mesure une compréhension raisonnée peut également être atteinte par une participation citoyenne à la recherche.

Cet article s'interroge ainsi sur les modalités de développement des sciences participatives en TALN. Il pose en particulier certaines questions qui se doivent d'être étudiées en regard de la nature particulière de son sujet d'étude, et afin de réfléchir au mieux à la part que l'on souhaite voir donnée au citoyen dans les orientations scientifiques :

- *Jusqu'où ?* – Différentes modalités d'interventions citoyennes en sciences peuvent être imaginées. Nous ferons tout d'abord un retour historique sur les rapports entre science et public pour montrer que le citoyen conserve ses capacités de compréhension en présence d'une science de plus en plus complexe et spécialisée. Partant de ce constat, nous discuterons ensuite des différentes gradations d'interventions qui peuvent être envisagées. Nous en conclurons qu'un apport citoyen doit être recherché avant tout en amont (définition de questions scientifiques et apport d'observations) et en aval (évaluation des technologies) du processus de recherche.

- *Comment ?* – Nous nous interrogerons sur les modalités déjà effectives en TALN et à la manière de conduire une recherche participative dans cette discipline. Nous ferons d'une part le recensement de travaux existants dans le domaine, puis nous proposerons avant tout un cadre méthodologique pour étudier la modalité et l'apport d'une recherche participative en TALN. Il nous semble en effet que notre discipline manque encore de

[7] Voir http://www.allianceathena.fr/c/sciences-participatives/.

recul sur la question pour permettre l'émergence d'une science citoyenne au-delà des discours convenus.

- *Mais pourquoi ? Retour sur un TALN participatif* – Nous montrerons enfin en quoi les sciences participatives peuvent présenter un intérêt pour le TALN. Nous reviendrons tout d'abord sur la nécessité éthique d'un regard citoyen sur la direction de nos recherches, mais nous suggérons également que les sciences participatives peuvent être un moteur éventuel d'innovation par le regard décalé qu'elles apportent.

2 Jusqu'où : quelle implication citoyenne ?

2.1 Science et compréhension citoyenne : mise au point historique

Des échanges plus marqués entre public et scientifiques peuvent favoriser le développement du progrès technique. Cet échange n'est toutefois concevable que si le citoyen est à même de saisir en profondeur les enjeux des recherches concernées. En présence d'une science hyper spécialisée, certains doutent qu'une réelle intercompréhension puisse encore être atteinte et s'accordent sur une séparation désormais indépassable entre chercheurs et citoyens. Avant d'étudier plus en avant les frontières d'une recherche participative en TALN, nous nous appuierons sur une analyse historique des rapports entre science et opinion publique (Bensaude-Vincent 2013) pour indiquer que l'existence de ce fossé infranchissable n'est pas démontrée.

La science moderne, envisagée comme l'établissement de vérités par une analyse critique des résultats d'épreuves de vérification et de falsification d'hypothèses, émerge au XVII^{e} siècle et s'établit réellement dans l'esprit des Lumières au XVIII^{e} siècle. Cette cristallisation de la pratique scientifique est concomitante avec l'émergence d'un vrai espace politique (cafés, salons, loges maçonniques) pour l'opinion publique. Dépositaire de l'autorité de l'analyse critique, le scientifique y a toute sa place, et en retour le public est lui aussi invité à participer, comme observateur et expérimentateur, au développement de la science. Bien qu'elles soient ouvertes à tous, ces interventions citoyennes ne concernent qu'un public lettré. (Habermas 1978) parle ainsi d'*espace bourgeois*. Il n'en reste pas moins que la science recourt alors à ces amateurs éclairés pour se constituer (voire se financer). Un des mérites de l'Université sera d'étendre ultérieurement ce public éclairé à d'autres catégories sociales.

Le XIX^{e} va correspondre à une institutionnalisation de la recherche scientifique soutenue par l'opinion publique. L'essor de la presse scientifique, concomitant à la Révolution Industrielle, répond à une curiosité croissante des masses pour la recherche, considérée alors comme le principal facteur de progrès de l'humanité. Ainsi renforcée, la communauté scientifique va alors cantonner le public dans un rôle passif de soutien : la vulgarisation n'est plus considérée que comme un outil politique pour assoir la promotion de la recherche. Ainsi, l'Académie des Sciences interdit elle, à la mort (1853) de François Arago, secrétaire perpétuel et grand vulgarisateur, les interventions du public en séance. La séparation d'avec le public qui s'observe à ce moment historique n'est donc pas due aux progrès rapides de la science, mais à une stratégie de disqualification de la part d'une communauté scientifique qui cherche à se créer un espace social (Stenghers 1993 : 183-185).

Cette rupture sera totalement consommée chez Bachelard au XX^{e} siècle. S'appuyant sur le mythe (Latour & Woolgar 1988) fondateur d'une recherche rationnelle objective, la science devient chez

cet épistémologue très influent une nouvelle religion à laquelle le public doit se soumettre aveuglément (Carnino 2015) : la parole de l'expert ne peut plus être contestée, l'opinion publique est totalement délégitimée (Bachelard 1938). L'opinion procède pour Bachelard uniquement de l'émotion et non de la rationalité permise par l'exercice de la raison critique chez le scientifique. C'est cette autorité de la raison critique qui justifie l'abandon au seul chercheur de la définition de ses directions de recherche. Malgré certaines réserves exprimées[8], une telle position ne saurait être affirmée aussi clairement désormais.

Il faudra attendre l'émergence de la société du risque technologique (Beck 2001) et la répétition de scandales écornant l'image du scientifique rationnel indépendant de toute pression extérieure (climato-sceptiques, expériences sur l'innocuité du tabac ou de l'amiante financées par l'industrie), pour que l'opinion publique ose à nouveau s'intéresser à la conduite de la science, à une période où apparaissent précisément les sciences participatives (années 1970). Des initiatives associatives ou institutionnelles (*Main à la pâte*, *Sciences en Fête*) cherchent également à combler ce fossé à partir de la fin du XX^{e} siècle, sans interroger par contre le rôle respectif des chercheurs et du citoyen. Ce dernier pas semble toutefois franchi comme en témoigne le soutien affirmé par le CNRS avec sa mission « Sciences et citoyens » (Wolton 2013).

Par ce bref retour en arrière historique, nous avons cherché à montrer que la disqualification de l'expertise citoyenne n'est pas due à un quelconque analphabétisme scientifique mais à une stratégie d'isolement de la communauté des chercheurs à un moment historique de son histoire. Rien ne justifie le maintien d'une telle rupture, qui est plus ou moins prononcée suivant les disciplines et ne semble plus répondre à un programme de disqualification de l'opinion publique, comme ce fut le cas à l'époque de Bachelard. Reste à déterminer la place que peut prendre le public dans ses échanges avec la communauté scientifique, au-delà de certains discours incantatoires convenus. Nous allons aborder cette question en étudiant dans un premier temps les différentes modalités d'action qui ont été envisagées pour la mise en place d'une science citoyenne.

2.2 Quelles modalités d'action pour une science citoyenne

Dans sa recommandation sur les sciences citoyennes, le comité éthique COMETS du CNRS distingue quatre formes d'action participative : le recueil d'information, la science distribuée où l'interprétation complète l'observation des données, la co-conception et enfin la « science citoyenne extrême » où l'amateur vise une réelle contribution théorique (COMETS 2015). Nous préférons trois catégories différentes d'action qui peuvent potentiellement concerner le TALN.

Soutien passif à la science : *crowdfunding* – Ici, le citoyen contribue au développement de la science sans chercher nécessairement une compréhension profonde du fait scientifique. Nous regroupons dans cette catégorie le *crowdfunding*, où le particulier oriente ses dons vers un programme de recherche (*Téléthon* par exemple) ou sur des plateformes de financement participatif destinées à la recherche[9] telles qu'il en existe dans le monde anglo-saxon. On peut également citer le calcul scientifique distribué sur les ordinateurs de particuliers consentants dans des domaines

[8] Voir la prise de position de l'Association Française pour l'Information Scientifique sur la légitimité du chercheur à juger seul de ses orientations de recherche : www.pseudo-sciences.org/spip.php?article2501.

[9] Voir par exemple les plateformes *SciFund Challenge* (https://scifundchallenge.org) et *Experiment* (https://experiment.com/) dans le monde anglo-saxon. Des initiatives équivalentes commencent désormais à voir le jour en France, comme avec la plateforme DaVinciCrowd (http://www.davincicrowd.com) de l'IFFFRES (Institut Français des Fondations de Recherche et d'Enseignement Supérieur).

comme la génomique ou l'astronomie[10]. Cette forme de soutien traduit un intérêt pour la science, mais ne garantit pas d'un effort profond de compréhension critique. Le risque est ainsi de voir cet investissement orienté vers des champs scientifiques aux enjeux plus visibles, ou faisant appel à l'émotion, dans le domaine de la santé en particulier. Le co-pilotage de la recherche cherche au contraire à éviter ce type d'écueil en visant une réflexion éclairée de la part du citoyen.

Participation au pilotage de la recherche : *conception centrée utilisateur* – L'intégration d'un avis citoyen dans le pilotage de la recherche requiert une intercompréhension entre chercheurs et grands publics sur les enjeux des travaux visés. C'est une démarche dont l'intérêt est reconnu des institutions, comme le montrent les débats nationaux menés sur des sujets comme les OGM et les nanotechnologies (CNDP 2010). Le caractère houleux de ces consultations montre la difficulté de la mise en place d'une implication citoyenne à ce niveau de décision si celle-ci n'est pas réfléchie en amont. Les expériences en sciences participatives, telles les conventions de citoyens élaborées par la Fondation Sciences Citoyennes (Testard 2015), montrent toutefois qu'en suivant certaines bonnes pratiques, une telle co-conception de la recherche est parfaitement atteignable. Pour cela, il faut viser une co-construction de la question scientifique à aborder, opérer une médiation/formation qui permette à tous les acteurs de comprendre les enjeux concernés, et penser la gouvernance pour que les citoyens acteurs ne soient pas dépossédés de la conduite du projet (Houlier & Meridhou-Goudard, 2016). La mise en place de panel d'utilisateurs volontaires pour juger de l'acceptabilité de certaines technologies (en IHM ou en aide au handicap en particulier), peut conduire à une forme faible de co-pilotage si les sujets recrutés ne sont pas considérés uniquement comme sources d'observation de comportements, mais sont au contraire associés à tout le cycle de développement logiciel.

Apport de données scientifiques : *crowdsourcing* – La collecte de données scientifiques est une forme répandue de science participative. Elle a perduré de toute date, et de manière essentielle, dans des sciences de l'observation telles que la zoologie et la botanique[11], l'astronomie[12], la météorologie mais également l'archéologie. Cette type de participation, appelé *crowdsourcing*, se retrouve en TALN pour la production de ressources linguistiques. Elle peut y prendre des formes parfois originales, telles que par exemple la participation du public à des jeux sérieux : outre la plate-forme *JeuxdeMots* , déjà évoquée, ou peut citer les exemples de *Phrase Detectives* (Chamberlain et al. 2009), jeu coopératif en ligne permettant l'annotation d'anaphores ou encore *Zombilingo* (Fort *et al.* 2014) qui permet de participer à l'annotation d'un corpus en dépendances syntaxiques. Dans ces trois exemples, le joueur est bénévole, ce qui peut nous interroger sur la possible externalisation d'une recherche scientifique non financée. A l'opposé, les recherches en TALN ou en humanités numériques ont de plus en plus recours à la plateforme Amazon Mechanical Turk (AMT) pour rémunérer des particuliers (les *turkers*) à la tâche afin de produire des données d'observation. Cette alternative pose des problèmes de contournement du droit du travail, de même qu'il a été montré qu'elle ne se traduisait pas par une amélioration de la qualité de données (Sagot *et al.* 2011). La collecte citoyenne de données rémunérée doit donc être limitée par des contraintes éthiques (COMETS 2015), de même que l'on peut s'interroger sur la nature de la participation mobilisée dans ce cas : retrouve-t-on l'engagement citoyen attendu des sciences participatives ?

---

10 Voir par exemple le programme *SETI@home* de calcul distribué pour la recherche de vie intelligente extraterrestre : http://setiathome.berkeley.edu/.

11 Voir par exemple en France, dans ce domaine, le rôle moteur du Muséum National d'Histoire Naturelle mais aussi de sociétés naturalistes locales.

12 Voir par exemple les différents programmes de repérages de cratères lunaires, de classification de galaxies proposés sur la plate-forme de sciences participatives Zooniverse : www.zooniverse.org.

Le *crowdsourcing* est intéressant en ce sens qu'il permet d'étudier le niveau d'expertise que peut atteindre un public motivé dans un domaine spécialisé ou, à minima sur une tâche donnée. Comme l'ont montré plusieurs projets suivant une démarche participative (Fort 2016), les données obtenues par *crowdsourcing* peuvent répondre aux exigences de qualité du TALN du moment où la formation des individus est soutenue ou que la cohorte des participants est de taille suffisante[13]. Le degré d'expertise atteint par *crowdsourcing* et la nature de l'engagement du citoyen semble très variable entre un naturaliste amateur et un participant d'un jeu sérieux. Mais ces différentes formes de participation témoignent toutes d'un intérêt salutaire pour le fait de science. Le *crowdsourcing* ne dégage toutefois pas le chercheur de l'activité de récolte d'observations puisque lui revient la définition du protocole de collecte et de validation des données obtenues.

2.3 Co-construction de connaissance

Le citoyen impliqué dans une démarche active de sciences participatives peut donc atteindre un niveau certain d'expertise. Se pose alors la question de savoir jusqu'où il est à même de participer à la co-construction de la connaissance scientifique. Au XVIII^{e} siècle, certains amateurs atteignaient une maîtrise expérimentale qui faisait d'eux de vrais acteurs de la science. En règle générale, ils ne faisaient toutefois que reproduire les protocoles expérimentaux définis par un savant[14]. Qu'en est-il aujourd'hui ?

A notre connaissance, rares sont les exemples de science participative où le public se place au niveau de l'interprétation de données, et encore moins de la production de théories ou de modèles qui sont au centre de l'activité du chercheur. Certes, les projets PICRIS de la région Ile de France sont systématiquement co-pilotés par un laboratoire de recherche et une association. Mais cette co-conception se focalise sur la définition des objectifs de recherche et l'évaluation des résultats, deux lieux d'intervention naturels pour un contrôle politique de la science. Alors que l'expertise du chercheur reste essentielle pour la définition des protocoles de collecte de données, leur interprétation, et la construction des modèles qu'elles permettent d'établir.

Certains promoteurs des sciences participatives, telle la *Fondation Sciences Citoyenne*[15] défendent une conception plus poussée de la co-conception des connaissances. A notre connaissance, aucun projet de cette nature n'a été mené en TALN. Les seuls exemples d'une implication citoyenne sur l'ensemble du processus de recherche relèvent d'un activisme politique méfiant vis-à-vis d'une science officielle. C'est le cas de groupes de contre-expertise tels que la CRIIRAD dans le domaine du nucléaire. On peut également citer la mouvance des *hacklabs* ou *biohackerspaces*. Ceux-ci peuvent viser des actions d'appropriation ludique de la recherche – (Meyer et al. 2012) parle de

---

13 Certains joueurs assidus de *ZombiLingo* atteignent ainsi un niveau de compétence élevé (précision proche de 90%) sur une tâche d'annotation syntaxique (Fort 2016:90). Karën Fort parle d'experts de la tâche (et non pas d'experts en linguistique) tout en insistant sur le fait que, comme pour toute acquisition de compétence, la formation est essentielle (Fort 2017). L'absence de formation obligatoire sur AMT peut expliquer les études montrant que les experts produisaient des données de meilleure qualité que la foule de *turkers* (Bhardwaj et al. 2010 : section 6). A l'opposé, plusieurs études montrent qu'une cohorte d'annotateurs naïfs formés peut égaler la qualité d'annotation d'un expert isolé sur certaines tâches (Wilbur 1998 ; Zesch & Gurevych 2009 ; Lafourcade & Joubert 2013).

14 À l'exception du Longitude Prize, issu du Longitude Act (1714), loi du parlement britannique accordant une somme considérable à quiconque déterminerait par une methode simple la longitude à laquelle se situait un navire en mer. Le Longitude Prize actuel propose un challenge autour de la résistance aux antibiotiques (https://longitudeprize.org/).

15 Fédération Sciences Citoyennes : http://sciencescitoyennes.org/.

bricolage – mais aussi atteindre une « science citoyenne extrême » (COMETS, 2015) qui se développe en marge de tout cadre institutionnel (création d'un *hackerspace* dédié à un pan scientifique non exploré par les laboratoires académiques par exemple), et ignore parfois les contraintes éthiques qui encadrent la recherche scientifique (Meyer 2012, Fiévet, 2015). Comme les amateurs du XVIIIe, ces *biohackers* acquièrent une maîtrise expérimentale réelle. Mais parle-t-on encore de science lorsque ces travaux parfois exotiques sont menés sans contrôle éthique ? Est-on encore en présence d'une recherche citoyenne irriguant l'ensemble de la société ?

Nous ne le pensons pas : ces exemples de « science citoyenne extrême », nous rappellent simplement l'existence de chercheurs indépendants aux côtés de la recherche institutionnelle. L'exemple de *Bretagne Vivante*, illustre cette situation : créée en 1959 par des militants écologistes sous le nom de *Société pour l'Etude et la Protection de la Nature en Bretagne* (*SEPNB*), elle a développé progressivement une expertise qui l'a conduit à une institutionnalisation l'écartant de la science citoyenne ordinaire : elle salarie des chercheurs, parfois d'anciens bénévoles mais surtout des étudiants formés par l'Université, qui participant de manière classique à la production de connaissances avec des collègues universitaires. Sa particularité reste de combiner une action militante dont la légitimité est reconnue des pouvoir publics à des activités conventionnelles de recherche et de conservation du littoral. Ceci en associant étroitement bénévoles associatifs et chercheurs, suivant un modèle citoyen participatif qui pourrait irriguer la recherche publique.

De fait, les chercheurs de *Bretagne Vivante* sont des scientifiques qui n'ont pas oublié qu'ils étaient des citoyens : ayant recours à la participation du public pour la collecte d'information voire leur interprétation, leur pratique scientifique reste classique. Elle reste par contre grande ouverte à un échange citoyen au niveau de la direction et de l'évaluation des programmes de recherche. Cette vision de la co-conception de la connaissance scientifique est, de notre point de vue, la seule opérative. Nous aimerions toutefois nous interroger sur ses modalités de sa mise en œuvre en TALN, faute d'un recul épistémologique et expérientiel de la communauté sur le sujet.

3 Comment : sciences participatives en TALN

Pour envisager ce que pourrait être une science participative approfondie en TALN, nous proposons de considérer les principales étapes de production de connaissance du domaine :

Production d'observables – On regroupe ici la constitution de corpus annotés, de ressources lexicales, d'ontologies essentielles au développement des technologies langagières. Une telle participation citoyenne relève du *crowdsourcing* dont nous avons déjà discuté le bien fondé.

Interprétation des données – A notre connaissance, aucune recherche participative en TALN n'a fait intervenir à ce jour le citoyen dans l'interprétation des données qu'il produit. Pourtant, des plateformes comme *ZombiLingo* et *JeuxdeMots* disposent de forums où s'expriment certaines interrogations linguistiques. Par exemple, Nobody, participant à *ZombiLingo* demande :

> *j'ai eu un énoncé avec le verbe dépendent* [sic] *et un groupe sujet avec coordinations de deux GN, la réponse était la tête du GN1… Je me demandent* [sic] *ce que vont valoir les annotations*

Le fait de ne pouvoir faire porter la relation sujet sur la coordination en elle-même semble gêner le joueur. Indirectement, il interroge le schéma d'annotation en dépendances retenu par le jeu, par rapport à une annotation en constituants sans doute plus proche de ses connaissances scolaires. Ce

type d'échange nous amène à penser que le public pourrait aider le chercheur à finaliser certaines conventions d'annotation, pour les rendre plus opératives. Cette participation a sans doute ses limites et peut scléroser le débat scientifique : on pense ainsi à la gestion chronophage de sollicitations de pertinences variables qu'elle pourrait entraîner. Mais il nous semble intéressant d'étudier expérimentalement la capacité de l'amateur à participer à de tels échanges, sans avoir accès aux modèles théoriques manipulés par le chercheur. L'objectif étant, une fois encore, d'apporter un point de vue original, ancré dans l'expérience naturelle de locuteur de l'amateur.

Développement de systèmes – La création des technologies langagières requiert au moins deux expertises complémentaires : une connaissance sur les modèles de langage intégrés dans les systèmes, et une compétence informatique pour leur développement. Il est difficile d'imaginer qu'un public non expert puisse mobiliser cette compétence pluridisciplinaire pointue. L'existence du mouvement *open source*, composé de contributeurs particulièrement motivés, invite toutefois à considérer avec attention la contribution citoyenne qui pourrait être attendue dans ce domaine : détection de *bugs*, retours d'expériences, réutilisation dans des applications hôtes, ajouts de plugins. L'intervention citoyenne concerne ici l'expression de besoins et également l'appropriation avancée des technologies. De nombreux systèmes de recherche sont désormais diffusés en open source. Le système *Sibylle* d'aide à la communication pour personnes handicapées (Wandmacher *et al.* 2008) est diffusé librement par le centre de rééducation de Kerpape dans sa version *Sibylle vK*[16]. L'analyse des téléchargements montre que le logiciel est utilisé par de simples particuliers qui n'hésitent pas à faire des retours sur l'utilisation de dernier. Le site de diffusion de *Sibylle* fait par ailleurs appel à la contribution bénévole de plugins utiles à la communauté.

Evaluation – L'évaluation devrait être un des champs d'intervention privilégiés des sciences participatives en TALN, dont les applications commencent à se déployer sur des données parfois sensibles : fouille d'opinion et monitoring sur les réseaux sociaux, identification d'auteurs (Statamatos 2009) par exemple. La participation citoyenne à l'évaluation des systèmes participe donc au principe éthique de contrôle citoyen du progrès technologique. Elle présente deux intérêts pour notre communauté scientifique. D'une part, elle permettrait de remettre en contexte une vision parfois mythique des possibilités du TALN et plus généralement de l'Intelligence Artificielle, telle que dispensée par certains chercheurs eux-mêmes dans les médias. Cela éviterait les déceptions qui ont déjà affecté l'histoire du TALN par le passé, mais surtout écarterait l'application inconsidérée de nos technologies sur des sujets tels que, par exemple, l'expertise judiciaire (Boé *et al.* 1999).

D'autre part, le TALN ne cesse, à raison, de s'interroger sur ses pratiques d'évaluation. Des campagnes d'évaluation concernant les modèles d'évaluation eux-mêmes existent même dans des domaines tels que la traduction automatique (Macháček & Bojar 2013). Un regard citoyen est susceptible d'aider le chercheur à se poser certaines bonnes questions en matière d'évaluation ; par exemple en se focalisant sur des métriques liées à l'acceptabilité des technologies et plus facilement compréhensibles pour le grand public. Il ne faut bien entendu pas négliger les difficultés à trouver un panel de citoyens motivés par une participation à des actions d'évaluation. Ainsi que l'a montré l'aide au handicap, on pourrait toutefois imaginer que des associations représentatives soient associées à certaines évaluations. De même, l'implication d'informaticiens appelés, au-delà du seul secteur des industries de la langue, à utiliser nos technologies, serait enrichissant.

[16] Sibylle vK : http://k-lab.fr/sibylle/.

4 Mais pourquoi ? retour sociétal sur un TALN participatif

4.1 Nécessité éthique : science en conscience et société du risque

Nous avons ainsi cherché à montrer qu'une implication citoyenne plus forte peut être opérative en TALN. Dans cette perspective, il semble important de s'interroger sur les motivations profondes qui doivent présider à une telle évolution. Un premier argument en faveur de l'inclusion du public dans le débat scientifique est d'ordre politique. L'époque (pas si lointaine au regard de l'Histoire) du chercheur ou de l'inventeur indépendant est révolue, et une part significative de la recherche est réalisée sur financement public. L'activité du chercheur dépend donc du soutien que veut bien lui accorder la communauté des citoyens. D'un point de vue pragmatique, il ne peut éviter un regard citoyen critique sur les enjeux et les directions de sa recherche. La responsabilité de la recherche scientifique soulève toutefois des questions éthiques qui ont reçu des réponses diverses au fil du temps : cette responsabilité doit-elle être confiée à la société civile au nom de laquelle sont menées ces recherches, aux chercheurs qui dispose de l'autorité scientifique, ou bien cette direction doit-elle être l'objet d'une consensus sociétal éclairé ?

Cette question de responsabilité éthique touche tous les aspects de la production scientifique. On la retrouve dans la controverse sur l'édition scientifique payante. Outre le fait qu'elle délègue à des intérêts privés des résultats issus le plus souvent de la recherche publique, l'édition payante écarte *de facto* le public de l'accès direct aux idées et résultats scientifiques. Le citoyen doit alors s'en remettre au filtre des médias de vulgarisation. A l'heure d'Internet, le mouvement *open access* aux publications, mais aussi l'existence de supports pédagogiques en ligne de plus en plus pointus sont des démarches qui doivent être rapprochées des sciences citoyennes, même s'ils elles ne seront pas discutées ici. De même, le mouvement en faveur d'une libération des données et des modèles[17] apparaît comme un prérequis à une réelle démocratie technique (Callon *et al.* 2001). Elle se retrouve dans le débat actuel sur la transparence des algorithmes[18], qui a donné lieu à un rapport auprès du Ministère de l'Industrie et à une consultation actuellement en cours sous l'égide de la CNIL[19]. Cette problématique citoyenne est très visible dans le cas des algorithmes utilisés par la finance à haute fréquence. Mais ce serait une erreur que d'ignorer que le TALN est également concerné alors qu'apparaissent des algorithmes de fouille de texte ou de recommandation travaillant automatiquement sur des données textuelles[20].

Pourtant, une intervention citoyenne plus marquée dans la conduite de la recherche ne doit pas être vue comme une contrainte mais au contraire comme une forme de libération pour le chercheur. La dimension éthique de sa responsabilité sociale peut en effet être lourde à porter isolément dans notre société du risque technique (Beck 2001). Un regard citoyen permet un partage de

---

17 Voir par exemple http://internetactu.blog.lemonde.fr/2014/12/26/ouvrir-les-algorithmes-pour-comprendre-et-ameliorer-les-traitements-dont-nous-sommes-lobjet/.

18 Voir par exemple la controverse à propos des algorithmes de la plateforme Admission Post-Bac : http:/www.lemonde.fr/campus/article/2017/03/02/apb-nouvelle-etape-vers-la-transparence-de-l-algorithme_5088173_4401467.html. Par ailleurs, à la suite de la loi Lemaire, INRIA a lancé TransAlgo, une plateforme « pour le développement de la transparence et de la responsabilité des systèmes algorithmiques, du fait de la dualité des données et des algorithmes » (https://www.inria.fr/actualite/actualites-inria/transalgo).

19 Débat éthique et numérique : les algorithmes en question (https://www.cnil.fr/fr/ethique-et-numerique-les-algorithmes-en-debat-0).

20 Pour une illustration qui croise TALN et trading à haute fréquence, voir http://www.france24.com/fr/20130424-ap-twitter-hack-piratage-explosion-maison-blanche-effet-bourse-wall-street-high-frequency-trading.

responsabilités sur les conséquences de ses recherches, fardeau qui se fera de plus en plus pressant : l'exemple de la condamnation de sismologues suite au tremblement de terre de L'Aquila est là pour le rappeler (Grazzi 2012). Or, le TALN arrive à un stade de développement où ses applications ont des implications économiques, sociologiques et éthiques importantes que nous ne saurions négliger (Lefeuvre *et al.* 2015).

4.2 Regard décalé et innovation scientifique

Les bénéfices que l'on peut retirer des sciences participatives ne sont pas que défensifs. En particulier, un regard d'usager expert peut être un moteur stimulant d'inventivité scientifique. En dépit de sa démarche objective de vérification et de falsification des preuves, on sait que la science académique est susceptible de tomber dans un dogmatisme freinant toute innovation (Kuhn 1983). Face à ce risque, le recours à une science citoyenne peut parfois être un atout.

Prenons l'exemple de la plateforme participative *JeuxdeMots* (Lafourcade et Joubert 2008) et de ses multiples petits jeux associés, où des particuliers aident en s'amusant à la construction d'un réseau lexical. Analysant les contributions reçues, les auteurs de *JeuxdeMots* observent que les réponses des experts linguistiques sont « *correctes mais peu imaginatives et n'alimentent pas la longue traîne* », c'est-à-dire les relations pertinentes les plus rares, contrairement aux contributions du public (Lafourcade & Joubert 2013 :208). Ainsi, le regard décalé du non expert peut-il contribuer, dans le cas présent, à la collecte d'observations originales. Les *a priori* dogmatiques, la facilité d'une pensée conservatrice ne sont bien sûr pas l'apanage du chercheur académique. Mais la confrontation à une connaissance ou des besoins différents ne peut être qu'un stimulant pour la réflexion scientifique. (Bensaude-Vincent 2013) distingue ainsi deux formes de savoirs mobilisés par les acteurs d'une recherche participative :

- pour le public, un savoir local (au sens où il est contextualisé) qui relève d'une recherche permanente d'adaptation guidée par l'expérience personnelle,
- pour le chercheur, un savoir global nourri de la pratique de vérification et falsification des hypothèses, et d'un effort de modélisation et formalisation qui lui confère son universalité.

C'est la confrontation de ces deux types de savoirs qui est enrichissante pour le chercheur impliqué dans une démarche de recherche participative. On peut ainsi espérer du citoyen actif la (re)formulation de questions qu'il conviendra au scientifique de remettre ensuite en perspective.

5 Conclusion : pour une évaluation des sciences citoyennes

En conclusion, cet article ne doit pas être considéré comme un plaidoyer en faveur des sciences participatives en TALN, mais comme une première réflexion sur ses modalités d'action. De fait, l'existence des sciences citoyennes dans la recherche publique est déjà une réalité, et elle est actuellement encouragée par les pouvoirs publics, comme le montre par exemple la récente mission « Sciences participatives » (Houlier & Meridhou-Goudard, 2016). Alain Fuchs, directeur du CNRS, voit même dans l'émergence d'une science participative forte un moyen de ré-acculturation scientifique de l'opinion publique face à la montée en puissance de discours pseudo-scientifiques (Dessibourg 2017). Notre propos ici a donc été de montrer en quoi les sciences participatives, finalement elles-aussi déjà assez présentes dans le domaine du TALN sans qu'elles ne donnent leur nom, sont un apport non négligeable aux avancées de celui-ci. Analysant les différentes activités liées au développement des technologies langagières, nous avons esquissé des champs

d'intervention citoyenne qui vont au-delà de la seule participation à la collecte de données (*crowdsourcing*). Nous avons en particulier insisté sur l'intérêt d'une association du public à l'évaluation des technologies, et nous sommes interrogés sur la limite que l'on peut définir à la formation du public par une pratique active à la recherche. À notre connaissance, ces questions n'ont pas reçu de réponse dans le domaine du TALN. Aussi pensons-nous qu'il serait intéressant de mener des études expérimentales portant sur l'apport citoyen dans des projets de sciences participatives. L'existence d'une science participative forte dans un domaine de recherche nécessite l'implication de citoyens motivés. Cette motivation peut être essentiellement ludique dans le cadre des jeux sérieux. La nature de cette motivation pourrait certainement être diversifiée, du fait de l'intérêt que tout un chacun porte à sa langue maternelle. Une implication plus forte, que ce soit en terme de formation et de désir de participation au pilotage de projets de recherche, demande sans doute que le citoyen perçoive un enjeu politique derrière les travaux réalisés. Nul doute que l'apparition somme tout récente d'applications du TALN posant des enjeux éthiques visibles renforcera ce besoin.

Remerciements

Les auteurs tiennent à remercier les relecteurs de l'article pour leurs commentaires. Ils leur ont permis d'approfondir leur réflexion sur les modalités d'approfondissement des sciences citoyennes en TALN. Nous tenons également à remercier Pierre Halftermeyer pour les échanges fructueux auxquels il a contribué sur l'intérêt et les limites d'une participation citoyenne à la recherche scientifique.

Références

ANTOINE J.Y., VILLANEAU J., LEFEUVRE A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. Proc. *14th Conference of the European Chapter of the Association of Computational Linguistics, EACL'2014, Gothenburg, Suède*.

BACHELARD G. (1938) *La Formation de l'esprit scientifique*. Contribution à une psychanalyse de la connaissance objective, Paris : Vrin.

BHARDWAJ V., PASSONNEAU R., SALLEB-AOUISSI A., IDE N. (2010) Anveshan: a framework for analysis of multiple annotators' labeling behavior. Proc. *4th Linguistic Annotation Workshop* (LAW IV), Uppsala, Suède, 47-55.

BECK U.(2001). *La société du risque*, Paris : Flammarion, Champs/essais

BENSAUDE-VINCENT B. (2013). *L'opinion publique et la science : à chacun son ignorance*. Paris : La découverte.

BOE J.L., BIMBOT F., BONASTRE J.F., DUPONT P. (1999) Des évaluations des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. *Cahiers d'Etudes et de Recherches Francophones. Langues*, 2 (4), 270-288.

CNDP, COMMISSION NATIONALE DU DEPART PUBLIC (2010) Bilan du débat public sur le développement et la régulation des nanotechnologies. Avril 2010.

CALLON M., LASCOUMES P., BARTHE Y. (2001) *Agir dans un monde incertain. Essai sur la démocratie technique*. Paris : Seuil.

CARNINO G. (2015) *L'invention de la science : la nouvelle religion de l'art industriel*. Paris : Seuil. L'univers historique.

CHAMBERLAIN J., POESIO M., KRUSCHWITZ U. (2008) Phrase Detectives : a web-based collaborative annotation game. Proc. *Int. Conference on Semantic Systems, I-Semantics'2008*.

COMETS (2015) *Avis du COMETS : les sciences citoyennes*. Rapport comité COMETS du CNRS.

DESSIBOURG (2017) *Les chercheurs debouts face à Trump*. La Recherche, 522, 16 :19.

FIEVET C. (2015) Vive les apprentis sorciers. *Uzbek & Rica*, 13, 92-98.

FORT K., GUILLAUME B., CHASTANT H. (2014) Creating ZombiLingo, a Game With A Purpose for dependency syntax annotation *Proc. of the Gamification for Information Retrieval (GamifIR'14) Workshop*. Amsterdam, Pays-Bas.

FORT K. (2016) Collaborative annotation for reliable natural language processing. Lindon : ISTE, Wiley. Coll. Focus, Cognitive Science Series.

FORT K. (2017) Experts ou (foule de) non-experts : la question de l'expertise des annotateurs vue de la myriadisation. Revue Corela, volume HS-21 « Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage » (http://corela.revues.org/4835).

GADET F. (2006) La Variation sociale en français. Nouvelle édition revue et augmentée. Paris : Ophrys

GRAZZI F. (2012) Quelle expertise scientifique après le verdict du procès de L'Aquila ? *La Recherche*, 470, décembre 2012.

GUILLAUME B., FORT K., LEFEBVRE H. (2016) Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax *Proc. COLING'2016,* Osaka, Japon.

HOULLIER F, MERILHOU-GOUDARD J-B. (2016). Les sciences participatives en France : état des lieux, bonnes pratiques et recommandations. Rapport à la demande des ministres de l'Education nationale, de l'Enseignement Supérieur et de la Recherche. Février 2016.

HABERMAS J. (1978) L'espace public, archéologie de la publicité comme dimension constitutive de la société bourgeoise. Paris : Payot.

KUHN T. (1983) *La structure des révolutions scientifiques*. Paris : Flammarion coll. (ed. révisée).

LAFOURCADE M., JOUBERT M. (2008) Jeux de Mots : un prototype ludique pour l'émergence de relations entre termes. Actes *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*.

LAFOURCADE M., JOUBERT M. (2013) Bénéfices et limites de l'acquisition lexicale dans l'expérience Jeux de Mots *In*. GALIA N., ZOCK M. *Ressources lexicales. Contenu, construction, utilisation, évaluation.* John Benjamins Publ.

LATOUR B., WOOLGAR S. (1988) *La vie de laboratoire : la production des faits scientifiques*. Paris : La Découverte. (trad. ed. anglaise 1979).

LEFEUVRE-HALFTERMEYER A., ANTOINE J.-Y., COUILLAULT A., SCHANG E., ABOUDA L., SAVARY A., MAUREL D., ESHKOL-TARAVELLA I., BATTISTELLI D. (2016) Covering various Needs in Temporal Annotation: a Proposal of Extension of ISO TimeML that Preserves Upward Compatibility. Proc. *LREC'2016*. Portoroz, Slovenia.

LEFEUVRE A., ANTOINE J.-Y., ALLEGRE W. (2015) Ethique conséquentialiste et traitement automatique des langues: une typologie de facteurs de risques adaptée aux technologies langagières. Actes de l'atelier *Ethique et TRaitement Automatique des Langues* (*ETeRNAL'2015*), *conférence TALN'2015*, Juin 2015, Caen, France, 53-66.

MACHÁČEK M., BOJAR O. (2013) Results of the WMT13 Metrics Shared Task. Actes *ACL'2013 eigth workshop on statistical machine translation, WMT'2103*. Sofia.

MEYER M. (2012) Bricoler, domestiquer et contourner la science : l'essor de la biologie de garage. *Réseaux*, 173, 303:328. http://www.cairn.info/article.php?ID_ARTICLE=RES_173_0303.

NORVAL C., HENDERSON T. (2017) Contextual consent: ethical mining of social media for health research. in *Proceedings of the WSDM 2017 Workshop on Mining Online Health Reports.* WSDM Workshop on Mining Online Health Reports, Cambridge, United Kingdom, 10-10 February.

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. Actes *TALN'2011.*

STATAMATOS E. (2009) A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538-556.

STENGHERS I. (1993) *L'invention des sciences modernes*. Paris : La Découverte.

TESTARD J. (2015) *L'Humanitude au pouvoir : comment les citoyens peuvent décider du bien commun*. Paris : Seuil.

WANDMACHER T., ANTOINE J.-Y., DEPARTE J.-P., POIRIER F. (2008) SIBYLLE, an assistive communication system adapting to the context and its user. *ACM Transactions on Accessible Computing*. 1(1). 1-30.

WILBUR W. J. (1998). A comparison of group and individual performance among subject experts and untrained workers at the document retrieval task. *JASIS*, 49(6), 517–529.

WOLTON D. (2013). Légitimons le débat entre scientifiques et citoyens. JOURNAL DU CNRS, 272, mai-juin 2013.