-à paraître dans Lapurdum 2013 - http://lapurdum.revues.org/

La linguistique comme science ouverte
Une expérience de recherche citoyenne à carnets ouverts sur la grammaire du breton

Mélanie Jouitteau
IKER, UMR 5478

Meilleur qu'hier, pire que demain

La pratique de la science est en 2013 à l'aube précoce d'un bouleversement profond et irréversible, déclenché par le média internet et ses possibilités nouvelles. Dans la littérature anglophone, la diversité des domaines dans lesquels des expériences de science ouverte sont menés est grandissante (voir la webliographie de Dawson 2012). Dans cet article, je présente dans une première partie ce que je comprends de la science ouverte en 2013, et des possibilités nouvelles que cette pratique offre à mon champ de recherche en particulier, la linguistique formelle. J'opère dans une seconde partie un retour sur pratique à partir de mon expérience de science « à carnet ouvert ». Je développe depuis 2009 un centre de ressource et recherche en ligne de type wiki, dans le cadre d'une recherche en syntaxe formelle. Le site de plus de 2000 pages comprend en particulier une grammaire descriptive et formelle de la langue bretonne ouverte à contributions et discussions. La grammaire comprend en avril 2013 près de 500 articles, et l'ensemble comptabilise plus de trente visites par jour en moyenne sur l'année 2012. L'interactivité y est naissante et prometteuse. J'explique comment ce média transforme mes pratiques de recherche et comment l'accès large au processus et aux résultats de cette recherche transforme sa « réception ». J'expose les possibilités que je vois dans la formule de recherche à carnets ouverts. Je conclus sur les conditions d'émergence d'une science ouverte et citoyenne. [i]

1. Recherche fermée, recherche ouverte

La recherche telle que le XX° siècle et le début du XXI° siècle connaissent est une recherche fermée en ce que sa valeur motrice provenait de la restriction de l'accès aux savoirs. Un article ou un livre à succès est un ouvrage qui, à au moins un lieu donné, peut voir son accès tarifé. Le profit généré alimente les organismes de diffusion de la recherche (librairies, maisons d'édition, plates-formes de distribution, etc.), qui en retour déploient un arsenal juridique destiné à préserver la restriction d'accès aux savoirs (copyrights, contrats de cession des droits à l'éditeur). Ce prix est donc un "prix du douanier", puisque c'est directement l'impossibilité construite d'accéder à l'ouvrage par des moyens alternatifs qui transforme la valeur scientifique de l'ouvrage en une valeur chiffrée, et assure *in fine* sa diffusion (relative). La recherche fondamentale, peu susceptible d'être alimentée par le secteur privé, est toute entière structurée par ce "prix du douanier". La recherche appliquée l'est par le système des brevets. Individuellement, les actrices de la recherche sont récompensées pour leur capacité à produire des ouvrages dont la diffusion sera restreinte, afin de pouvoir monnayer leur accès. Entre autres inconvénients, ce modèle demande au secteur public de la société de payer deux fois pour profiter des avancées de la science: une première fois par le biais du financement public de la recherche, et une seconde fois pour pouvoir accéder aux résultats potentiels de cette recherche.

L'enjeu sociétal de l'accessibilité des résultats des recherches en sciences est très important. La science fermée est, ce n'est pas nouveau, une science de reconduction des rapports de classe, puisque seules les classes monétairement et culturellement privilégiées possèdent à la fois les moyens de payer l'accès aux produits de la recherche, et les moyens culturels d'en faire un investissement culturel ou financier qui leur rapporte. Ce qui est nouveau, c'est que la justification du coût de reproduction du livre sur lequel ce système est basé est en train de tomber. Une personne atteinte d'une maladie grave ou inconnue pouvait comprendre que le coût de reproduction des ouvrages scientifiques à faible distribution l'empêche de pouvoir se renseigner elle-même sur ce qui met sa vie en danger. Cette même personne a maintenant naturellement le réflexe de chercher accès aux recherches en cours. Elle trouve des plateformes de distribution des publications tarifiées à des prix prohibitifs, alors que son expérience concrète de multiplications de fichiers et de formules de partages gratuits lui démontre que ces produits seraient aisément reproductibles. Dans mon champ empirique qui est celui de la langue bretonne, langue associée à un stigmate de classe assez fort, écrire pour les accédant·e·s à la recherche fermée revient à exclure plus de 99% des personnes pratiquant la langue des fruits de la recherche, dans un contexte de demande sociétale forte, au moins en terme de description de la langue. [ii]

1.1. Un bouleversement avec précédents

L'histoire humaine a connu avant le XX° siècle deux bouleversements majeurs de la diffusion des savoirs. L'invention de l'écriture a permis d'atemporaliser la passation des savoirs: nous pouvons maintenant passer des savoirs d'un humain à un autre sans que ces deux humains ne se rencontrent, ni même ne partagent le même espace-temps. L'invention de l'imprimerie a ensuite permis de démultiplier drastiquement la diffusion des écrits, permettant à un humain de diffuser un savoir vers des milliers d'autres, à travers l'espace et le temps, et ce à un coût de production bien moindre. L'invention d'internet et des logiciels de partage tels qu'ils apparaissent à l'aube du XXI° siècle intervient comme un troisième bouleversement majeur. Ce nouveau modèle de création, agrégation et diffusion des savoirs transforme profondément l'histoire des savoirs humains, et ce depuis une dizaine d'années seulement.

La question de savoir si ce changement fondamental dans la donne est un bien ou un mal envahit les médias, mais est, je pense, de peu d'intérêt pour les linguistes. En tant que scientifiques, au pire, nous assistons à la naissance de la recherche ouverte. Au mieux, nous y participons. Je pense pouvoir montrer que le gain pour l'avancée des sciences, et pour le coeur de nos métiers, même en sciences humaines, peut être considérable.

1.2. Que recouvre le terme de science ouverte?

La science ouverte est construite dans l'esprit du logiciel libre, et assume cette référence. Sa forme la plus interactive est appelée *science 2.0.*, en référence au passage du web du XX° siècle, immense mais figé, distribuant des informations fixes à une communauté élargie, au web du XXI° siècle, encore plus vaste et transversal, média essentiellement nouveau qui fonde une culture d'échanges de contenus transformables. Un système d'exploitation libre comme *Linux* a

pu être développé par une communauté souple qui a su inventer des solutions de développement collaboratif à grande échelle car son code source est libre d'accès. La science ouverte s'inscrit dans le même modèle de développement que le logiciel libre, et utilise les mêmes outils. Via le réseau mondial internet, l'objet scientifique est développé de façon ouverte et transparente, avec ses données et sa méthodologie entièrement accessibles en ligne à lecture, commentaire et transformation. La science ouverte favorise l'utilisation des techniques nouvelles pour inventer des efficacités collaboratives. Les formes compétitives ne sont pas obligatoirement évacuées, mais les compétitions prototypiquement deviennent des formes de compétitions entre des formes collaboratives. La collaboration y est cumulative, car les outils numériques permettent une agrégation massive des résultats. Elle est aussi distributive car la collaboration prend forme en tâches différenciées : des intervenantes différentes peuvent prendre en charge des formes d'aide de différente ampleur, distribuant ainsi la tâche collaborative en des formes souples adaptées aux propositions (Nielsen 2012).

2. Exemples pionniers de science ouverte

La propriété définitoire de la science ouverte est l'accès libre aux résultats de cette recherche, mais d'autres propriétés de cette pratique émergent, comme les pratiques de crowdsourcing actifs et passifs, les sondages en ligne, les projets cartographiques de masses de données, et les pratiques de recherche à carnets ouverts. Je les discute et illustre brièvement dans cette partie.

2.1. Accès libre

Au début des années 2000, des premiers pas vers l'accessibilité des publications scientifiques ont été faits et sont maintenant relativement bien installés culturellement. Les articles scientifiques, ou tout du moins leur pré-version avant publication, sont de plus en plus disponibles en ligne. En linguistique dans le domaine anglophone, certains modèles émergent de la communauté à l'initiative d'un individu, comme dans le cas de *Lingbuzz* développé à l'université de Tromsø par Michal Starke depuis 2004. Il n'est pas anodin que cette initiative soit venue d'un scientifique avec une double formation à la linguistique formelle et à la programmation informatique. Cette double formation réunit deux prérequis majeurs: la capacité technique de réaliser un programme en ligne, et la culture du logiciel libre.

Dans le domaine francophone, des plateformes naissent par le biais institutionnel, comme HAL ou Artxiker développé au sein du laboratoire IKER. Il n'existe à ma connaissance aucune plateforme sur le web où seraient regroupées les ressources numériques disponibles pour l'étude de la langue bretonne.[iii]

2.2. Crowdsourcing et restitution naissante

Le crowdsourcing désigne l'action de chercher des sources de données dans la multitude. Ce terme est importé directement du vocabulaire des économistes. Il est introduit par Howe (2006), à la suite de Surowiecki (2004, 2008), pour cerner ce que Surowiecki appela la « sagesse de la foule », le 'plus' produit par la collaboration de masse rendue possible par les outils naissants. Ce terme n'a pas à ce jour de traduction satisfaisante en français dans le domaine scientifique. Le

terme d'*externalisation ouverte* est un terme managérial désignant une forme de sous-traitance, et le terme de *collaborat* n'a pas la transparence et la spécificité du terme anglais. Quelques sciences traditionnelles ont une pratique installée de crowdsourcing. Les sciences de la nature font ainsi depuis longtemps appel au grand public pour recueillir des observations simples mais multiples de la faune et de la flore. On peut voir depuis une décade cette culture évoluer vers la restitution des résultats, voire, mais ce n'est encore qu'une propriété émergente, vers la restitution des données brutes vers le grand public. Je montre ici à travers trois exemples de projets concernant la biologie, *Allo Elasmo* (2013), *Project squirrel* (1997-2013), *Project BudBurst* (2007-2013) comment s'installe progressivement l'idée de restituer les données vers la source multiple.

En 1997, le *Project squirrel*, lancé par Wendy Jackson et Joel Brown, a fait appel au grand public afin d'observer les écureuils fauves et les écureuils gris à Chicago. En quinze ans, un millier de participantes ont envoyé des observations ou photos par mail à une équipe scientifique. Cette équipe a ensuite posté les données sur le site et mis en ligne un article publié à partir des données (Van der Merwe, Brown et Jackson 2005). Le délai de restitution est considérable (8 ans), et les données brutes ne sont pas accessibles. Ce modèle est toujours en cours dans l'Etat français : la campagne d'Apecs (2013), *Allo Elasmo*, conduite avec le Parc naturel marin d'Iroise, fait appel aux usagers de la mer par des campagnes d'affichage et via internet, pour reporter leurs observations de raies et de requins en mer d'Iroise. Un guide d'identification des espèces est téléchargeable en ligne afin d'augmenter la fiabilité des observations recueillies. Les informations sont récupérées par téléphone, synthétisées puis intégrées comme données dans des travaux scientifiques éventuels. Le retour aux contributrices est assimilable à un travail de vulgarisation classique : une fois le travail scientifique présenté en interne à un public scientifique, un rapport final est posté en ligne et consultable par les internautes. Le *Project BudBurst*, aux Etats-Unis, depuis 2007, collecte des données sur les effets du changement climatique en appelant les internautes à noter pour des plantes communes en Amérique du Nord les dates et coordonnées géographiques d'apparition de premières feuilles, floraison, fructification, etc. Les données sont envoyées en ligne sous forme numérisée. Comme dans les modèles précédents, une équipe de scientifiques poste en ligne un rapport annuel en pdf, mais les données sont aussi intégrées dans des outils cartographiques *Googlemap* accessibles en ligne. La banque de données soutenant la carte n'est cependant pas téléchargeable. La restitution des données brutes sous une forme visuelle est cependant un pas nouveau. Elle permet par exemple une utilisation personnelle des données: une personne allergique aux pollens d'une fleur particulière peut prévoir une médication appropriée lors d'un voyage. Le fonctionnement est encore loin d'une banque de données téléchargeable en ligne que tout scientifique ou citoyen pourrait utiliser librement pour tester des corrélations.

Le mouvement évolutif va vers l'instantané avec des délais de restitution des données qui se raccourcissent. Une dimension nouvelle émerge, qui tend à rendre utilisables les données brutes par des actrices internes comme externes à la communauté scientifique. Dans le domaine de la linguistique formelle, quelques initiatives émergent qui ont cette propriété nouvelle de

comprendre des données brutes accessibles en ligne. A ma connaissance, ils ne sont pas (encore) ouverts à participation en dehors du monde scientifique. Un bel exemple est SSWL, *Syntactic Structures of the World's Languages* (2009-2013), un site qui construit une base de données en ligne des propriétés linguistiques des langues du monde. Sur invitation, plusieurs centaines de linguistes du monde entier sont appelés à typifier leurs langues d'expertise, en déposant des exemples emblématiques pour chaque structure grammaticale proposée à comparaison. Un tutoriel vidéo permet de se familiariser rapidement avec l'interface, qui, encore peu attractive et intuitive, est visiblement destinée à un public de spécialistes. L'outil est cependant incontestablement d'une puissance toute nouvelle, avec un potentiel immense. Il grandit régulièrement : la base de données comprend plus de 176 langues, avec plus de 71 propriétés linguistiques associées. Il est d'ores et déjà possible d'obtenir en ligne une carte des langues qui ont deux propriétés croisées, et de consulter les phrases exemples associées avec leurs gloses. Le potentiel d'un tel projet pèse d'autant plus lourd si on considère que ce projet est sans financement depuis 2010, et se développe depuis entièrement en termes de bénévolat d'experts. [iv]

2.3. Crowdsourcing et traitabilité des données

Certaines données brutes ne sont pas exploitables telles quelles, et demandent un traitement préalable. Le site Galaxy Zoo fait appel à des volontaires pour classer en ligne les milliers de photos prises par un télescope robotique. Les questions posées à l'internaute sont typiquement simples pour un humain mais très dures à automatiser (*Voyez-vous une galaxie en spirale ou en ellipse ? S'il s'agit d'une spirale, tourne-t-elle dans le sens des aiguilles d'une montre ?*). Depuis 2007, plus de 150 millions d'images de galaxies qui pour la plupart n'avaient jamais été vues par un œil humain auparavant ont été classifiées par plus de 200.000 volontaires en ligne (Nielsen 2012 :129). Le processus de classification a permis de découvrir, souvent par le biais de participant·e·s sans aucune formation particulière, l'hypothèse de l'existence de miroirs de quasars dans l'univers, ou de « galaxies en petits pois », des types de galaxies jusqu'ici inconnues (Nielsen 2012 :140).

En théorie, tout domaine dans lequel une masse lourde de données peut être traitée par une multitude d'individus sans que cela requière de formation trop importante pourrait trouver des formules de traitement des données via le crowdsourcing. La tâche, et ce n'est pas le plus mince des défis, doit juste être décomposable en microtâches et apparaître attrayante pour les internautes. Un autre projet impressionnant d'ingéniosité fut d'ailleurs *Foldit* (Popović & Baker 2007-2013), un logiciel écrit en forme de jeu où les internautes trouvent les meilleurs moyens de plier une protéine. Les résultats de meilleurs 'pliages' sont directement considérés comme données de recherche par les biochimistes (Nielsen 2012 :146).

Dans le domaine des langues minoritaires en Europe, des formules sont à portée de main. Il existe partout, des masses conséquentes d'enregistrements, dans les archives des radios locales, numérisées ou non, qui ne sont pas même répertoriées. Avec une interface appropriée, le crowdsourcing pourrait être utilisé pour transcrire ces documents. Dans une formule alternative semi-institutionnelle, pour le breton par exemple, la transcription en ligne pourrait être envisagée

par les étudiant·e·s des départements de breton des universités de Rennes et Brest. Cet exercice pédagogique supervisé pourrait fournir instantanément un corpus transcrit directement reversable au matériel pédagogique accessible en ligne. Le corps enseignant, passé la phase d'adaptation à l'exercice, pourrait voir son travail allégé par les commentaires en lignes d'autres traductrices expérimentées, et l'émulation en ligne.

2.4. Élicitations en ligne et sondages wikis

Les sciences sociales, linguistique incluse, procèdent traditionnellement du haut vers le bas quand il s'agit d'obtenir des données considérables statistiquement (par questionnaires aux choix possibles prédéterminés), ou du bas vers le haut sur des groupes plus restreints, avec un coût de traitement des données plus lourd (interviews, élicitations). Les médias de type wiki permettent d'inventer des questionnaires en ligne qui permettent des mouvements du bas vers le haut, avec des coûts de traitements moindres et un potentiel de réponses considérables statistiquement. Ces protocoles sont des « sondages wiki » (Salganik & Levy 2012), qui ont trois particularités saillantes : (i) ces interfaces ne rejettent pas de données a priori (aucune contribution n'est jugée trop petite ou trop grande), (ii) elles sont collaboratives et répercutent automatiquement le retour obtenu. Par exemple, au lieu de la case « autre réponse » opaque d'un sondage, la personne qui répond au sondage peut choisir de créer une nouvelle réponse. Celle-ci sera automatiquement intégrée aux choix proposés par le sondage, et sera visible pour la prochaine personne interrogée. Finalement (iii), le questionnaire est adaptable et modifiable au fil des réponses. Un tel modèle rendu possible par les techniques nouvelles propose un pas considérable dans la réduction de la tension entre recherche qualitative et recherche quantitative. Créer un de ces sondages est aisé et gratuit sur *allourideas.org*, qui développe le logiciel adapté et profite des retours sur fonctionnement pour améliorer le code et étudier les usages nouveaux qui émergent. Actuellement, le modèle développé par Salganik à l'université de Princeton ne permet encore ce type de sondage interactif qu'à partir d'une seule question donnée, ce qui restreint son usage pour la création de protocoles linguistiques en ligne, mais l'outil n'est pas loin d'être adapté aux besoins des linguistes.

Les protocoles d'élicitations sont rarement obtenus en ligne dans la communauté de linguistique générative, traditionnellement attachée à des séances d'élicitation en tête-à-tête, dont l'enregistrement audio même est rarement opéré et encore moins disponible. Dotlatčil & Nilsen (2009) fournissent un contre-exemple intéressant. Ces deux syntacticiens formels ont construit des protocoles en ligne pour tester des locutrices du néerlandais pour les types d'antécédents (universels distributifs, anti-additifs, pluriels définis, conjonctions de noms) qui pouvaient lier des anaphores réciproques et des prédicats distributifs. Le protocole en ligne a connu un grand succès, avec 1150 réponses rapides. Ces réponses ont constitué un petit ensemble traitable statistiquement, rendant négligeable la possibilité de réponses-plaisanteries qui auraient faussé les résultats. Atout non-négligeable, les réponses, opérées en lignes, arrivaient aux chercheurs entièrement numérisées, offrant ainsi une relative facilité de traitement. Ces exemples sont amenés à se multiplier d'autant plus que la manipulation des logiciels sera possible pour des linguistes sans formation de programmation.

2.5. Cartographie et organisation de masses de données

Les nouvelles échelles de masses de données accessibles demandent de nouvelles façons d'organiser ces données. La cartographie est une de ces solutions, et de nombreux projets ont une dimension cartographique. Ainsi le *Allen Brain Atlas* (2004-2013), qui cartographie le cerveau jusqu'au niveau des cellules individuelles, ou le *Ocean Observatories Initiative*, qui développe des outils pour fournir des informations sur le fond des océans en temps réels et en accès libre. Des projets similaires existent pour la mise en carte de l'univers, de l'haplotype ou du génome humain. Dans le domaine de la linguistique formelle aussi, la possibilité nouvelle de croiser des banques de données complexes commence à permettre au réflexe cartographique de s'incarner.

Le projet *Edisyn*, financé par la Fondation Européenne pour la Science (ESF) a essayé, à l'échelle européenne, d'initier et regrouper des banques de données linguistiques de langues européennes diverses et de les rendre cross-interrogeables en ligne, avec le développement d'outils en accès libre pour la visualisation cartographique de recherches croisées. Ce projet a été financé de 2005 à 2010, avec une extension partielle jusqu'à mars 2012. La base de données commune en ligne en 2013 croise des données de microvariation syntaxique de l'italien, du portugais, de l'estonien, de l'anglais, du néerlandais, du slovène et d'un ensemble de dialectes scandinaves. Cette réalisation est réellement grandiose, même si en l'état les données brutes n'y ont pas encore trouvé leur solution pour être livrées de manière automatiquement transparente pour le grand public. Il me semble cependant évident que l'avenir de la recherche publique est contenu en germe dans ces projets, tant l'enjeu global est la façon dont les contribuables peuvent appréhender nos services.

2.6. Science à carnets ouverts

La science dite à « carnets ouverts » ou *open notebook science*, est une pratique qui requiert de rendre accessible le matériel brut de recherche dès sa collecte. Cette pratique va frontalement contre les cultures scientifiques installées car cela implique de laisser à voir les parties du matériel jugées inutilisables, ou même dénuées de sens. Dans le champ théorique, cela implique de laisser à voir les erreurs, les hypothèses en cul de sac, ou basées sur des prémices mal construits par lesquelles nous passons pour arriver finalement à une découverte. D'aucun se demandera quel est l'intérêt de publier des non-sens alors qu'il est possible de publier un travail abouti, édité, lisible et … fini. Ce qu'il faut considérer, c'est la possibilité actuelle de fournir en ligne un tel ouvrage fini, mais accompagné de tout l'historique de sa genèse. Le travail scientifique n'est ainsi pas livré comme une révélation, mais comme un travail scientifique transparent dans sa mise en œuvre. Au-delà du critère de transparence, la recherche à carnets ouverts permet aussi d'enrichir le processus de recension. Un ouvrage en ligne peut être relu et commenté à différentes étapes de sa conception, et ce par un public, au choix connu ou anonyme, restreint ou entièrement ouvert.

L'exemple prototypique de succès de science à carnets ouverts en sciences fondamentales est le *Polymath project* (2009). Tim Gowers (Cambridge University) a posté en ligne, sur un blog ouvert, un problème ardu de mathématiques pour lequel il espérait trouver une solution, le

théorème de Hales-Jewett sur la densité des objets tridimensionnels. 37 jours et 800 commentaires après, qui mélangèrent dans une même discussion ouverte des intervenant·e·s spécialisé·e·s mais à de différents degrés d'expertise, Gowers annonça qu'il pensait que le problème initial avait été résolu, et que la recherche avait même incidemment résolu un problème plus vaste. Ce projet en a entrainé d'autres et a jeté durablement en mathématiques les fondations d'une culture de la collaboration (Nielsen 2012 :1-3 ; 209-13).

Le projet ARBRES présenté ci-dessous est un exemple de recherche à carnets ouverts. C'est un centre en ligne de ressources pour l'étude de la syntaxe du breton et de ses dialectes. J'y développe une grammaire descriptive et formelle des dialectes du breton avec une dimension typologique comparative.

3. ARBRES, plateforme de recherche syntaxique à carnets ouverts

Le projet ARBRES est né en 2009 lorsque j'ai opéré en Bretagne une série de consultations auprès de travailleuses de la langue. J'ai interrogé, parmi mes connaissances ou lors de rendez-vous plus formels, des libraires, des enseignant·e·s, des membres de l'office de la langue bretonne. Je cherchais alors à étudier la faisabilité d'une collecte à grande échelle sur le territoire parlant, et étais en recherche de partenaires de collectes et collaboration potentiels. Puisque mon entrée au CNRS me permettait d'installer des projets à long terme, dans une perspective de service public, et puisque j'étais en train de rencontrer justement ce public, j'en ai profité pour leur demander, à chacun, ce qu'ils attendaient du monde de la recherche scientifique linguistique. Certaines réponses étaient non-conclusives, principalement car les personnes avaient du mal à se représenter qu'ils pouvaient poster des vœux à un service public, ou même considérer le CNRS en tant que tel. Certaines réponses pointaient vers des besoins que je n'étais pas à même de combler, comme par exemple le développement d'un GPS en breton, qui ressort du domaine de la linguistique appliquée. Certaines réponses, de manière récurrente, pointaient le manque de matériel de référence sur internet, tant pour l'apprentissage que comme ressource pour les locutrices. Fulup Jakez, de l'office de la langue bretonne, a suggéré précisément le développement d'une grammaire du breton en ligne.

Après cette série d'entretiens, j'ai demandé à l'informaticien de mon laboratoire d'alors de me livrer un site vide de type wiki, hébergé sur le serveur de l'institution. J'ai choisi une formule de site interactif pour permettre un potentiel collaboratif maximum. Pour la même raison, le site a été laissé ouvert au maximum. Le prérequis de participation est uniquement de se créer un compte sur le site. Seules les pages de description du projet sont protégées en écriture. Toutes les autres pages sont modifiables. A chaque page du site est associée une page de discussion, sur laquelle les utilisatrices peuvent aussi laisser des remarques ou questions.

J'ai d'abord mis en ligne des notes personnelles que j'avais accumulées depuis de nombreuses années, comme une bibliographie qui se veut exhaustive des travaux publiés sur la langue bretonne, et les traces de projets de recherche rédigés auparavant, comme la description d'une douzaine de phénomènes grammaticaux que je voulais étudier. Comme je voulais que ces projets très techniques soient accessibles en compréhension, j'ai petit à petit adossé ces fiches à

des articles d'une grammaire naissante. Lorsque ces fiches ont été trop nombreuses, j'ai regroupé leurs liens dans un index de la grammaire, ce qui à son tour a dessiné en creux des pans manquants de cette grammaire qui compte maintenant près de 500 articles, dont certains très fouillés, sur des points divers de la grammaire bretonne.

Au fur et à mesure que l'ouvrage en ligne émergeait, j'ai systématisé des séances de consultation personnelle avec des utilisatrices du site. Ces consultations m'ont permis de modifier le mode d'écriture du site, car je voyais ainsi comment le site était utilisable en réalité. Il est par exemple important de voir quel vocabulaire rebutait les lectrices non-spécialistes. Après une décade de formation en linguistique, j'avais oublié que des abréviations telles que 3SGM, pour « troisième personne du singulier masculin » pouvaient poser des problèmes de lecture des gloses. Je les évite maintenant au possible sur ARBRES. Une gamme de vocabulaire technique était inévitable. J'ai donc créé un glossaire cliquable, où les notions de base sont expliquées, et illustrées avec des exemples du breton. La création de ces fiches explicatives m'a permis de rendre cliquables tous les termes techniques pour lesquels je voulais mettre le lectorat à un clic d'une explication. Sur la suggestion d'un collègue d'IKER, Aritz Irurtzun, qui a vu que cet outil pouvait servir à des traductions anglais-français des termes techniques de notre champ, j'ai dupliqué les entrées du glossaire, et je les ai traduites en anglais sur une page séparée. Il est maintenant possible sur le site, à partir d'un terme technique en anglais, de trouver sa traduction en français, ce qui ouvre à des utilisations nouvelles.

D'autres problèmes qui se posent au lectorat sont semi-temporaires, car ils me semblent liés au monopole des « habitudes papier ». L'outil de feuilletage « un article au hasard » rebute ainsi parfois le lectorat actuel qui, s'il aime feuilleter un livre papier, n'aime pas se sentir perdu et ne jamais « pouvoir finir » l'ouvrage. Je ne peux qu'espérer que le meilleur atout de cet ouvrage sera que personne ne pourra vraiment jamais le « finir ». L'outil de recherche est aussi sous-utilisé, ce qui ne peut s'améliorer qu'en terme de visibilité de l'outil sur la page, et en attendant que le réflexe, inévitablement, se répande dans le lectorat. Lors de séances de surf accompagné, j'ai aussi pu constater que certaines utilisatrices n'utilisaient jamais l'outil de déroulement des pages. Il ne leur est accessible que ce qui est directement visible sur l'écran et elles ne peuvent jamais lire ce qu'il y a en dessous, puisqu'elles ne descendent jamais plus bas sur la page. Je n'ai pas de solution à ce dernier problème, mais je pense toujours maintenant à mettre un résumé éloquent en tout début d'article, ce qui est indépendamment recommandable.

Le site comprend maintenant une page d'accueil, un plan du site, une page de liens vers des sites concernant la langue ou son étude, et une page d'actualités où sont signalés les prochaines conférences ou évènements concernant la syntaxe du breton. Je décris aussi dans la page d'actualité les derniers travaux menés sur le site. Un lien vers les modifications récentes permet de voir en détail l'activité sur le site. La grammaire occupe une place importante, avec une page d'index, un glossaire des termes techniques en anglais et en français, une bibliographie générale, une liste de liens actifs d'abréviations utilisées dans le champ qui envoient à une explication ou une référence complète, une page très importante de remerciements et un outil de feuilletage « une page au hasard » qui distribue uniquement sur les pages d'articles de cette

grammaire. Le centre de ressources fournit une liste des locutrices bretonnes natives d'une variété dialectale spécifiée, ainsi que les corpus qu'ils ont produits. Ces corpus sont géographisés sur une carte *googlemap* qui permet de trouver, pour un fait grammatical décrit dans un ouvrage descriptif, quel corpus de variété dialectale géographiquement proche est susceptible de l'illustrer. Les types de corpus disponibles pour l'étude de la langue sont décrits sur une page à part, afin de faciliter la recherche pour des linguistes en recherche de corpus. La dernière réalisation en date est la centrale d'élicitation, où je mets en ligne le résultat des collectes opérées, parfois à la demande d'autres linguistes.

Durant le mois d'avril 2013, selon l'outil *google analytics* associé au site ARBRES, le site a reçu 1348 visites de 1118 internautes différentes, qui ont visualisé en tout 3960 pages. La moyenne par visite est de 2,94 pages, avec une durée moyenne de 2 :33 minutes. La fréquentation n'est pas associée à un moment particulier, si ce n'est une légère augmentation d'utilisation en semaine, et non le week-end. 78,5% des visiteuses effectuaient leur première visite sur le site, alors que 21,5% avaient déjà visité ce site auparavant. La plupart des utilisatrices ont un navigateur en français, mais tout de même plus de 12% utilisent un navigateur en anglais. La provenance des connexions dessine la carte des pays francophones excluant l'Afrique subsaharienne (France 62,76%, Canada 5,27%, Algérie 2,60%, Maroc 2,60%, Belgique 2,23%). Quelques pays de culture anglophone fournissent le gros des connexions restantes (USA 3,34%, Pays-Bas 2,08%, Royaume Uni 1,71%, Allemagne 1,56%). Les villes de l'Etat français dont proviennent le plus de connexions ne dessinent pas la carte de la zone extrême-Ouest de la Bretagne où le breton est traditionnellement parlé, mais la carte de la diaspora bretonne (Rennes 10,91%, Paris 10,31%, Quimper 2,82%, Brest 2,37%, Bordeaux 2,15%, Nantes 2%). On note sur ce mois d'avril, comme chaque mois, des épiphénomènes, comme 2,60% de connexions provenant de Tunis ou 1,63% de Levallois-Perret.

4. Particularités pratiques du média

Au fil des années, j'ai pu observer que la particularité de l'écrit en ligne a infléchi ma façon d'écrire. Le changement le plus attendu découle de la dimension "à carnets ouverts" de cette recherche. Je livre constamment, en ligne, observable par quiconque, un travail non-fini, susceptible de contenir des erreurs, orthographiques ou d'erreurs de raisonnement, qui ne me font pas honneur. N'importe qui, de façon anonyme ou non, peut (et devrait!) laisser un message sur le site relevant ces erreurs. Selon les façons d'y penser qu'on choisit, cela peut avoir une dimension proprement terrifiante. Cette dimension est aussi un bon fuel de travail, car il est possible de calmer cet inconfort quotidiennement par l'idée que le site est meilleur qu'hier et pire que demain. Cependant, les premières surprises passées, je ne trouve pas que cela fasse de la publication à carnets ouverts une publication fondamentalement différente du travail de publication de recherche en général. Recevoir des reviews a toujours un côté anxiogène. Nous voulons toujours réécrire l'article que nous avons écrit il y a cinq ans. La différence avec cet outil est finalement que les reviews arrivent de façon ininterrompue et que je peux réécrire l'article quand je le décide.

La transformation de l'écriture touche principalement deux points. Le premier tient au fait d'écrire sur un format de site web. Le second tient à la dimension participative du site. Je décris ici ces changements brièvement. Je ne rentre dans les aspects plus techniques de la façon d'écrire dans l'interface de modification d'un wiki que dans la mesure où cela éclaire les modifications sur le produit fini d'écriture.

4.1. Liens actifs

Une série de changements heureux est venue naturellement des liens actifs (cliquables). J'avais commencé en associant chaque article de la grammaire à une bibliographie particulière, en profitant de la non-restriction de l'espace que le wiki fournissait. Que la bibliographie soit longue en bas de chaque page n'est pas vraiment un handicap. Puis j'ai réalisé que toutes les références sur la page peuvent elles-même être cliquables, sans que le lectorat soit obligé de dérouler la page jusqu'en bas, ou qu'une note l'y emmène. Actuellement, une référence dans le texte telle que "Kervella (1947)" est associée à une page du même nom, qui regroupe les informations utiles concernant l'ouvrage. Ce type de pages comprend les références bibliographiques, mais aussi, selon ce qui est disponible, un résumé, une liste des extraits de cet ouvrage présents sur le site, les errata éventuels, l'historique d'édition, la typification du dialecte dont l'auteur est natif, les références de reviews de l'ouvrage par d'autres auteurs, ou même le lien direct au pdf de l'ouvrage en ligne. Ne restent actuellement sur la bibliographie de chaque article que la liste des ouvrages de recherche concernant le point grammatical précis de l'article. Ceci améliore sa lisibilité et restreint les redites à l'échelle du site.

L'utilisation des liens actifs permet aussi de faire vivre tout un réseau de significations dans le cœur même du texte. Lorsqu'une page nouvelle est créée, il lui est associé un titre, par exemple « les adjectifs ». Ce titre devient aussitôt une « adresse » vers laquelle un lien actif peut pointer : dans n'importe quel article du site, les mots « les adjectifs » peuvent devenir des liens actifs qui redirigent vers cette page. En pratique, il suffit dans le corps du script de mettre le terme entre double-crochets pour que le texte apparaisse cliquable sur le site : [[les adjectifs]]. L'utilisatrice qui clique sur ce lien sera dirigée automatiquement vers la page qui décrit les adjectifs. Si dans le corps du texte, le mot cliquable est « adjectival », le script sera alors [[les adjectifs|adjectival]]. Le script peut à première vue sembler barbare, mais j'ai trouvé simple et intuitif le langage wiki et je peux dire, comme les milliers de personnes contributrices aux wikipedias, qu'y entrer est relativement aisé.

Ce qui est vrai pour les titres d'articles l'est pour les fiches explicatives du glossaire. Tout terme technique, opaque pour le lectorat non-spécialiste, que je n'arrive pas à éviter est rendu cliquable. Ceci permet, dans une certaine mesure, d'élargir l'éventail du lectorat. Un terme non expliqué peut aussi être laissé en rouge par les internautes afin que j'y associe, plus tard, une fiche explicative. Contrairement aux sites comme *wikipedia*, la transformation du corps du texte de l'ouvrage en liens actifs n'est pas automatisée sur ARBRES. Cela signifie que j'opère moi-même le choix, suivant le contexte, de transformer en lien actif les suites de mots du site « les adjectifs ». Cette opération est couteuse en temps mais très payante en terme de lisibilité. Les

pages ne sont pas inutilement truffées de liens actifs, et le lectorat sait de confiance que le lien a été pensé pour son usage, et non opéré hors-contexte par un robot.

Les liens actifs permettent enfin un fonctionnement inattendu, que j'ai vu se développer avec l'émergence de la grammaire sur le site. Au fur et à mesure que plus d'éléments grammaticaux avaient une fiche qui leur correspondait, il est devenu possible de rendre les gloses de chaque exemple du site entièrement cliquable. Pour une personne qui apprend la langue, il est maintenant possible de naviguer d'exemple en exemple, en cliquant sur les éléments en glose incompris dans la phrase pour aller se renseigner sur cet élément. C'est une façon d'apprendre dont l'équivalent antérieur demandait de compulser des dictionnaires et grammaires divers, sans pouvoir être sûr de trouver l'article éclairant la notion nouvelle. Ce problème était d'autant plus épineux pour la langue bretonne que, sans connaître auparavant le système des mutations consonantiques propre à la langue et qui fait changer les mots sur leur consonne initiale, il était ardu et hardi de trouver ce mot dans un quelconque dictionnaire ou index, car on en ignorait justement avec certitude la première consonne. Par le biais des gloses cliquables, ce problème n'est plus.

4.2. Recherches et redirections automatiques

Dès sa création, le titre d'une page peut faire l'objet d'une recherche dans l'outil de recherche interne au site. Si l'utilisatrice tape « adjectifs » ou « les adjectifs » dans l'outil de recherche, la page associée apparaîtra. Cependant, à ce stade, si l'utilisatrice tape « adjectif » au singulier dans la boite de recherche, celle-ci s'avérera nulle tant que ne sera pas créée une page de redirection allant de « adjectif » à « adjectifs ». Les pages de redirection permettent de créer des réseaux de multiples pages menant à une même page centrale. Ainsi, si la page centrale apparaissant au lectorat est titrée « Les adjectifs », il est possible de créer différentes pages titrées « adjectifs », « adjectival », « adjectif », etc. qui vont chacune rediriger automatiquement la recherche vers la page centrale « les adjectifs ».

Ce système de redirection peut être utilisé pour dépasser des problèmes classiques de publication papier, comme les problèmes récurrents pour l'édition sur la langue bretonne dus à l'existence de multiples systèmes orthographiques. Par exemple, le verbe de situation [ema, emã] se trouve en corpus écrit sous la forme *emañ* en orthographe standard, mais aussi sous la forme *eman* ou *ema* dans des systèmes orthographico-dialectaux différents. La forme wiki permet de créer des liens actifs à partir de chacune de ces orthographes, avec un système de redirections qui pointent toutes vers une seule page. J'ai choisi que la page centrale soit elle-même en orthographe standard, mais ce choix est moins lourd de conséquences que dans un ouvrage papier, car la lectrice peut aboutir à cette page en cliquant sur une forme en corpus dans une différente orthographe, ou en tapant cette forme dans le moteur de recherche interne au site. Ce qui vaut pour différentes orthographes vaut pour différentes langues. Je peux décider qu'un terme technique dans un résumé d'ouvrage en anglais sur le site soit un lien actif qui mène vers sa traduction en français. Un mot de grammaire dans un titre d'article en breton peut tout aussi bien mener vers la page de l'étude de cet élément en français.

Par ailleurs, comme dans un traitement de texte normal, l'outil de recherche peut être utilisé pour retrouver une portion de texte en particulier. C'est très utile dans un site de plus de deux mille pages. Lorsqu'une utilisatrice du site corrige avec raison la traduction d'un des exemples en breton, je peux par exemple retrouver avec certitude toutes les occurrences de cette phrase et y répercuter la correction proposée.

4.3. Le média comme moyen d'accès

La particularité du média n'influe pas seulement sur le mode d'écriture qui y est développé. Il influe aussi, et c'est sans doute le plus intéressant, sur l'impact de ce qui y est développé. Le média permet ainsi la valorisation de l'accessibilité nouvelle des documents numériques. Pour exemple, la bibliothèque de l'université de Rennes II a un programme de numérisation de son fond qui, occasionnellement, met en ligne des documents en moyen breton, ou des documents pédagogiques anciens d'enseignement du breton (ou du français via le breton). Ces documents apparaissent référencés sur ARBRES dans les références de corpus ou dans la bibliographie générale. La dimension nouvelle tient dans le lien actif que j'ajoute à la ligne bibliographique, et qui mène instantanément au document intégral en ligne. Lorsqu'un exemple provenant d'un de ces ouvrages est mis sur ARBRES, le lectorat est donc littéralement à deux clics et un CTRL+F de la donnée dans son contexte original.

Ce travail participe à rapprocher la lectrice lambda des richesses immatérielles produites par le service public, et dans un esprit de création d'une culture de recherche, permet une proximité des sources qui appelle au réflexe de vérification des données. Certaines des interventions mineures de corrections de traduction sur ARBRES découlent vraisemblablement de cet accès facilité. Plusieurs fois, la lecture d'une anaphore a été rectifiée par une prise en compte du contexte que j'avais trop pauvrement opérée. L'accès direct à la source a permis aux relectrices de réaliser mon erreur et de la corriger.

4.4. Organisation et valorisation du heureux hasard

La littérature bretonne regorge d'ouvrages de microexpertise dialectale tels que : « Trésors parlés de <*localité*>… », « L'étude du breton de <*localité*> », « Recueil d'expressions et bons mots de <*localité*> », etc. Ces ouvrages sont souvent conduits par des locuteurs natifs, ou des experts d'un parler donné qui collectent dans leur entourage proche constitué de natifs. Ils contiennent des merveilles de données brutes pour l'étude de la microvariation syntaxique. Ces ouvrages, cependant, individuellement, ou en collection papier, peuvent s'avérer limités pour l'étude de la langue, à moins de posséder une mémoire infaillible et une capacité de synthèse peu commune. Feuilleter ces livres à la recherche d'une information précise est sans espoir. Sur le site ARBRES, j'agrège ces micro-connaissances ultraspécialisées en un ensemble comparativiste cohérent. En pratique, je parcours ces livres ainsi que des corpus divers, je relève les formes étonnantes et les organise en fournissant le contexte qui explique pourquoi ces données sont étonnantes. Ce travail amène en soi à développer de nouvelles généralisations intéressantes : la lectrice, in fine, qui s'interroge sur un sujet précis en trouvera une description détaillée et globale, enrichie de la description des variants et invariants. Ce travail pourrait être mené jusqu'ici par des moyens

traditionnels, mais la nature wiki du média amène ce travail plus loin en appelant et en valorisant la microexpertise en crowdsourcing.

Le pari de l'interactivité, c'est qu'à mesure qu'émerge la structure du site et qu'on se retrouve dans le paysage reconnaissable d'une grammaire théorico-descriptive, de nouvelles micro-connaissances ultraspécialisées pourront venir enrichir le contenu central. Sur une fiche descriptive complète des pronoms proclitiques, une locutrice donnée pourra ajouter une forme manquante, avec sa caractérisation dialectale et apporter ainsi une information qui, isolée, serait de peu d'intérêt, mais dans le contexte fourni, peut préciser la répartition dialectale d'une forme et éclairer par ricochet un phénomène entier dans une dimension diachronique, syntactique formelle, sémantique ou de phénomène de contact. L'interactivité encore timide sur le site porte déjà des fruits de ce type. Pour illustration, une des sources descriptives que je favorise sur le site pour sa clarté est la grammaire descriptive de Chalm (2008), auteur originaire de Cap Sizun. L. Jade, participant actif pionnier du site, lui-même capiste, a signalé une traduction d'exemple produite par Chalm qui montre un choix de traduction privilégiant le breton standard sur la version locale alternative.

(1) Kalz o deus kriet diwar o aon.
beaucoup 3PL a crié de leur peur
Beaucoup ont crié de peur. Breton standard, Chalm (2008:R.1.2.3)
Beaucoup ont pleuré de peur. Cap-Sizun, c.p. de L. Jade sur ARBRES, mai 2013

Les chances que j'aie moi-même accédé à cette information lexicale sur le dialecte du Cap et fait le lien avec l'exemple en (1) sont tout à fait négligeables. L'apport de L. Jade a été reporté le jour même sur la page bibliographique de Chalm (2008). Cet apport participe maintenant à évaluer la typification dialectale de l'ouvrage en entier, ce qui aura des conséquences importantes sur l'étude des phénomènes qui y sont décrits. L'outil wiki permet ainsi de faire fructifier la microexpertise en lui offrant l'écrin où elle devient utile et le réseau d'étude où elle est signifiante.

4.5. Crowdsourcing passif

Une dimension étonnante du média est la facilité d'obtention instantanée de statistiques d'utilisation. Bien compris, ces outils statistiques peuvent fournir une première forme d'interactivité avec le lectorat, une forme de crowdsourcing passif, fourni par les pages de statistiques du site wiki lui-même, et par un logiciel d'espionnage de google, tel *google analytics*.

Le tableau ci-dessous est le produit de relevés opérés sur la page publique de statistiques du site ARBRES. Depuis son ouverture en 2009, le site grandit à vitesse régulière, et comptabilise maintenant un cumul de plus d'un million de visites électroniques (humains et robots compris), ce qui traduit une réelle présence sur le web. En 2013, une recherche par mot clef concernant un item lexical de la langue bretonne ou bien un terme grammatical en français via les plus grands

moteurs de recherche internationaux donne régulièrement au moins une page de ARBRES dans sa première page de résultats.

(2) Données statistiques sur la page publique du wiki.

Date	Nombre global de pages	Nombre d'articles dans les pages	Visites sur le site	Utilisatrices enregistrées
20 avril 2009	139	59	14 553	27
23 avril 2010	423	366	69 616	34
21 avril 2011	813	609	324 757	47
07 janvier 2012	1046	738	554 189	68
05 déc. 2012	1570	964	879 600	238
02 février 2013	2082	1009	1 017 622	288
20 avril 2013	2162	1047	1 090 247	346

Les outils de comptage de *google analytics*, actifs sur le site depuis plusieurs années, sont plus précis et associés à un compte privé. Ils opèrent sur les visites « réelles » humaines, tout en excluant mes propres visites du comptage. Selon cet outil, le nombre de visites d'usagères, pour les quatre premiers mois de 2013 montre une moyenne de 1211 visites par mois, sans publicité particulière sur cette période.

4.6. Profil d'utilisation

Ce même outil statistique *google analytics* fournit une masse de données sur le profil d'utilisation du site. Il est par exemple possible de visualiser la liste des mots clefs utilisés sur les moteurs de recherche et qui ont abouti à une visite (ainsi que la durée moyenne de cette visite, si la visite a mené à d'autres pages du site, et si les utilisatrices avaient déjà visité le site auparavant). Je peux voir ainsi que depuis quatre ans, le nombre de visites induites par un mot clef item lexical breton a augmenté. Ceci est facilement explicable par l'intégration d'une masse toujours plus grande de données dans la langue. Ce type de retour statistique simple peut induire des comportements de réponse de ma part, puisque vouloir augmenter le lectorat peut amener directement à inclure une diversité toujours plus grande de données sur le site.

L'outil peut aussi permettre de cerner des demandes présentes dans le lectorat. J'ai pu par exemple constater en septembre 2009 que, sans doute suite au succès des ouvrages de Hervé Lossec sur les « bretonnismes » en français, une source de trafic mince mais régulière arrivait sur ARBRES suite à des recherches sur les particularités de ce français de Basse-Bretagne. Cela m'a incitée à regrouper dans un article dédié à ce dialecte du français toutes les notes comparatives avec le breton que j'avais disséminé au gré des articles sur le breton, dans les sous-parties « horizons comparatifs », au milieu d'autres comparaisons typologiques. Il est maintenant

possible de lire un descriptif grammatical assez complet des particularités du français de Basse-Bretagne, et d'aller voir d'un clic la description du phénomène comparable en breton, ce qui en fait un vrai outil pédagogique.

L'analyse des sources de trafic me permet aussi de constater, par exemple sur mars 2013, que la hausse de trafic depuis décembre est bien liée à l'établissement d'un glossaire des termes de la syntaxe formelle, puisque je vois apparaître dans la liste des recherches « grammaticalité des expressions idiomatiques », « grammaticalisation des adverbes », « question oui/non », « élimination traits ininterprétables syntaxe », « verbes inaccusatifs », « pronom impersonnel », etc. Cette tendance dessine un nouveau profil d'utilisatrices francophones, dont la recherche ne concerne pas initialement la langue bretonne. La durée de consultation des pages m'assure ensuite que la visite est conséquente. Le cas échéant, je peux, si les visites d'une page donnée sont trop rapides, aller m'assurer que la fiche est rédigée clairement et la modifier. Je peux voir aussi précisément si ces visites ont mené, après une explication de notion grammaticale illustrée par des données du breton, à la visite d'autres pages dans sur site.

Je peux améliorer presque en temps réel la visibilité du site, en étudiant ses usages. J'ai ainsi vu apparaître des visites de plus en plus importantes provenant de recherches par mots clefs de type « arbres Jouitteau » ou « arbres grammaire ». Ces recherches indiquent prototypiquement un usage prémédité de la ressource en tant que telle, et m'ont montré qu'un lectorat naissant se fidélisait. J'ai pu constater que l'adresse du site sur les forums de discussion apportaient un très mince flux vers ARBRES, alors que les adresses sur *wikipedia* en anglais ou en français amenaient un trafic plus conséquent. En consultant la liste des sites qui comprennent un lien actif vers ARBRES, j'ai découvert des forums que je ne connaissais pas, et qui discutaient de l'existence de la grammaire.

Ce type de retour en crowdsourcing passif permet de cerner grossièrement la satisfaction potentielle du lectorat du site. Le temps de visite des pages et les suites de pages visitées me donnent des indicatifs, même approximatifs, sur des zones potentielles de croissance : ARBRES a depuis sa création un taux de rebond assez fort, généralement au-dessus de 70%. Cela signifie que plus de 70% des visites sont des visites d'une seule page, c.à.d. que la lectrice a quitté le site dès sa page d'entrée sur le site. Si cette page d'entrée est la page d'accueil, il est probable que la visite soit un échec. Si la page d'entrée est une référence, il est probable que l'utilisatrice a trouvé ce qu'elle cherchait. Si la page unique est un article, il est probable qu'une réécriture pourrait amener plus de curiosité pour les autres pages.

4.7. Interactivité et crowdsourcing actif

La progression des utilisatrices enregistrées dans le tableau donné en (2) n'est malheureusement pas révélatrice d'une dimension participative florissante puisque la plupart des inscriptions sont des spams automatisés. Une veille quasi quotidienne du site me permet, manuellement, de bannir les comptes de ces utilisatrices factices, et éventuellement d'effacer les messages publicitaires qu'elles auraient déposés. En décembre 2012, une sécurité a été associée à l'ouverture de comptes, ce qui a réduit les inscriptions automatiques à un rythme plus gérable, mais toujours

conséquent, de 25 par mois. Les utilisatrices réelles sont encore rares, mais marquent une progression sensible. Comme dans les différents wikipédias, le gros des modifications opérées directement sur les pages sont le plus souvent d'ordre orthographique. Ces corrections sont faciles à gérer, car elles sont rarement injustifiées, et sur un ouvrage évolutif de plus de 2000 pages, cette aide à l'édition est plus que bienvenue. Viennent ensuite des commentaires, corrections ou questions qui sont déposés dans la page de discussion associée à un article donné. Ces interventions sont maintenant habituelles pour quelques intervenants systématiques. Le rythme de ce type de dépôts peut aller de zéro à une cinquantaine par semaine. Je réponds personnellement en ligne aux questions, et opère ou discute les changements suggérés. Jusqu'ici, toute intervention a reçu une réponse, au moins partielle, dans les trois jours. Je prévois que ce type d'interactions va se multiplier. Intervenir lorsque plusieurs personnes l'ont déjà fait et que l'on peut lire les réponses rapides (et, je l'espère, claires) postées en ligne devrait avoir un effet facilitant et désinhibant. Les intervenantes nouvelles potentielles peuvent aussi s'assurer, au vu des interactions précédentes, que le ton sur le site est celui du respect et de la construction collective d'une œuvre de bien commun, que les interventions sont valorisées et apportent souvent un mieux visible. Il est donc d'un enjeu certain pour l'ensemble du projet que chaque intervenante soit écoutée, respectée, et dûment remerciée. Une section entière du site est d'ailleurs dédiée aux remerciements à participation, remerciements qui incluent nommément (ou sous pseudonyme le cas échéant) les intervenantes sur le site, comme d'ailleurs les locutrices des élicitations et les personnes référentes que je consulte pour le développement du site.

Un dernier type d'intervention n'apparaît pas sur les comptages évoqués ci-dessus, mais a une grande importance. Je reçois en effet nombre de réactions spontanées au site par courriel séparé. Il s'agit majoritairement, en ordre décroissant, (i) de demandes de vérification de quelques sources citées dans les articles, transcription API, glose, traduction ou orthographe d'un exemple donné, (ii) de lettres d'encouragement et de félicitation pour le projet, dont certaines précisent comment le site a été porté à leur connaissance et l'usage précis qu'ils en font, (iii) de listes de propositions de corrections d'argumentaires, comme ce qu'il est possible de déduire d'un exemple dans un dialecte donné, ou l'interprétation d'un passage d'une grammaire descriptive. Ces sources de retour sur lecture *via* le courriel privé ont un profil particulier, car il s'agit pour l'instant exclusivement d'interventions venant du monde universitaire, qu'il soit professoral ou étudiant. Dans un cas unique, j'ai reçu une proposition spontanée de généralisation quant à la distribution d'un élément syntaxique (cf. article sur *nikun*, le mot négatif 'personne'). La généralisation proposée a été ensuite reportée sur la page de l'article correspondant, discutée et étayée sur le site, à partir d'une conversation épistolaire privée et de relevés de corpus contradictoires.

Finalement, et c'est une des réussites du site qui me semble importante et que je compte développer, je suis contactée par des linguistes en grammaire formelle générative qui demandent des éclaircissements ou des données nouvelles sur un point particulier du breton. Ce profil comprend étudiantes et chercheuses en Europe et au-delà. Pour citer les derniers exemples en date, Rachel Nye, une étudiante en syntaxe à l'université de Gand m'a contactée dans le cadre

d'une recherche sur la typologie des grammaticalisations de l'interrogatif de type 'comment' en complémenteur déclaratif. Le breton comprend effectivement un interrogatif de type 'comment', *penaos*, qui a des emplois déclaratifs émergents dans certains dialectes. J'ai mené une recherche descriptive ciblée sur le complémenteur *penaos* et ses usages déclaratifs et l'ai mis en ligne, en enrichissant dans ce sens la fiche « *penaos* » préexistante sur ARBRES. Le délai entre les premiers contacts et la livraison de la recherche ciblée a été de deux mois. Dans un second cas, j'ai été contactée par les chercheuses Nora Boneh (Hebrew University of Jerusalem) et Léa Nash (Paris 8), menant une collaboration de recherche sur l'usage des prépositions spatiales dans les structures verbales à trois arguments. J'ai demandé et reçu un protocole d'élicitation ciblé, développé par elles. J'ai traduit de l'anglais au français le matériel d'élicitation qui consistait en des tâches de traduction vers le breton. J'ai modifié légèrement le protocole initial en amont, pour éviter certaines complications attendues au vu de la structure du breton, et une seconde fois légèrement lorsqu'un mot particulier posait problème à la locutrice lors de l'élicitation. Le choix de la locutrice s'est fait suivant mes propres critères de facilité : je joue au théâtre avec la locutrice Yvonne Simon, qui s'est acquittée de la tâche rapidement et sans effort particulier – notre facilité d'échange préexistant à l'élicitation. Nous n'avions pas de salle privative lors du rendez-vous où enregistrer l'élicitation, et nous l'avons donc conduite (!) dans ma voiture. Cette « élicitation à la carte » a pu se faire dans un délai raisonnable : les premiers contacts ont été pris le 9 janvier, l'élicitation a été opérée le 23 mars, les réponses précises à l'élicitation ont été livrées en ligne le 25 mars. Le 27 mars, un relecteur du site a corrigé avec raison l'orthographe bretonne d'un prénom que j'avais orthographié à la française dans une phrase en breton (*Marie* > *Mari*). Sans l'outil particulier de recherche à carnets ouverts, il est plus que probable que les chercheuses n'auraient pas pu obtenir de réponses pour le breton. Auraient-elles voulu en obtenir qu'elles auraient fait face à de multiples difficultés : comment trouver des traducteurs fiables anglais/français/breton avec une pratique des protocoles de linguistique ? Comment trouver une locutrice pleinement bilingue correspondant à la tâche de traduction demandée (nous parlons d'une langue minorisée avec un fort stigmate social)? Comment choisir une « bonne » locutrice, c.-à-d. une locutrice native sans rupture de pratique, dont la souplesse d'esprit et l'écoute permettent cet exercice, et qui trouve raisonnablement plaisant de procéder à une tâche étrange et répétitive sans qu'aucune autre valorisation qu'un remerciement public soit offert ? Lors de cette élicitation pilote, le thème de recherche ne nécessitait pas le ciblage d'une locutrice d'un dialecte en particulier. Cette présélection peut cependant être cruciale pour d'autres thèmes de recherche, et peut faire partie du travail de « calibration » fourni à des projets de recherche via une centrale d'élicitation *wiki*.

4.8. Limitations

Pour qu'une expérience participative de science ouverte soit un succès, il faut impérativement entre les participantes une culture partagée des modes de discussion et des modes de raisonnement (Nielsen 2012 :chap5). Ce n'est pas automatiquement le cas quand partenaires scientifiques et non-scientifiques collaborent. Dans le cas de ARBRES, certaines contributions peinent à comprendre l'importance des règles de citation des sources, ou de vérifiabilité des

données, surtout lorsqu'il s'agit d'experts non-natifs qui peuvent prendre personnellement que leurs données soient mises en situation d'être discutées. Les contributions peuvent être par exemple teintées d'idéologies linguistiques diverses, de particularisme militant ou de territorialisme. Les exemples de réactions ci-dessous ne sont jamais apparus tels quels sur le site, mais chacun peut juger de leur vraisemblance et de l'impact qu'ils auraient sur un projet collaboratif :

(3)*Les locuteurs qui utilisent cette forme ont l'air d'idiots.*

(4)*Il (ne) faut absolument (pas) reporter les formes bretonnes qui ressemblent aux formes présentes en français !*

(5)*Ce qui n'existe pas dans mon dialecte n'existe nulle part, je le saurais !*

(6)*J'ai toutes les réponses aux questions que vous vous posez depuis longtemps, je tiens à le dire ici et n'en dirai pas plus car tout a été résolu il y a vingt ans dans un article que je distribue aux gens de confiance.*

Les réflexes illustrés ci-dessus sont bien connus des scientifiques pour la bonne raison que se former scientifiquement soi-même revient à apprendre à mettre ces réflexes, certes humains, de côté. Ces problèmes viennent d'une part dans la population d'un manque de familiarité avec la culture scientifique, et d'autre part d'un manque de culture du partage en sciences. Cependant, prendre le temps d'essayer de les résoudre prend une dimension réelle de vulgarisation qui va au-delà du principe de science ouverte. Je ne vois pas, à l'heure actuelle, et dans mon champ de recherche, de moyen d'échapper à ce préalable. Le succès final de ARBRES tiendra probablement à ma capacité potentielle à dénouer ces débats à partir d'une position d'ouverture.

5. Conclusion, économie du libre et conditions d'émergence

Comment se fait-il que la science ouverte ne soit pas plus développée, et quels sont les obstacles réels qui s'y opposent ? Il existe certes des freins administratifs, culturels et juridiques actuels au développement durable de la science ouverte, spécifiquement en France (se reporter à Farchy, Froissart & Méadel 2010, et à sa recension par Schöpfel 2011). Mais si ces barrières existent, quelles sont les conditions d'émergence, dans le paysage actuel, d'une généralisation des pratiques de science ouverte ?

Le système français et européen en général s'est infléchi nettement vers un financement privé de la recherche au moment historique où le modèle naissant du crowdsourcing permettait nouvellement aux industries de réduire drastiquement leur masse d'investissement dans la recherche. Howe consacrait en 2006 un quart de son article sur le succès nouveau du crowdsourcing au domaine des sciences. Il reportait l'expérience de *InnoCentive*, un site au départ pharmaceutique dédié à l'innovation technique et scientifique par crowdsourcing : l'idée fructueuse était de fournir une plateforme où les industries postent des problèmes scientifiques à résoudre avec un prix de récompense pour une personne ou équipe qui fournirait une réponse satisfaisante. Dans la masse des gens informés, des individus ont des parcours singuliers qui apportent rapidement des réponses sans investissement financier conséquent. Ce modèle donne

des résultats nettement positifs pour l'industrie marchande ou les ONG en butte à des problèmes techniques lors de projets de développement dans le Tiers monde (Nielsen 2012 :22-24). Howe notait dès 2006 qu'avec de tels modèles de financement de la recherche, l'industrie avait trouvé une opportunité nouvelle et efficace pour éviter d'investir le coût lourd de structures de recherche en interne. Ce moment est antérieur de plusieurs années au moment où, dans l'Etat français, le gouvernement a déclenché un bouleversement sans précédents de l'économie de la recherche avec une modification des sources de financement *vers* les laboratoires privés de l'industrie, avec une valorisation du désinvestissement financier par le secteur public. Le domaine du privé est par excellence le domaine du « prix du douanier », et les sciences appliquées, technologiques et pharmaceutiques en tête, supportent aujourd'hui ce handicap vis-à-vis de la science ouverte. Les sciences humaines et fondamentales, moins à même de trouver un financement dans le secteur privé, deviennent donc aujourd'hui les champs où la science ouverte est le plus susceptible d'émerger, et ce malgré un handicap culturel traditionnel de relation aux nouvelles technologies.

Le développement de la science ouverte, et par là de la science tout court tant il est évident que l'accélération du travail scientifique changera en profondeur ce que la science nous apporte, est actuellement portée par des scientifiques qui en sont récompensés uniquement par des modèles culturels valorisants. En conséquence, on voit naître des projets de science ouverte surtout dans le domaine informatique, ou l'histoire du libre a créé un système de valorisation symbolique puissant. Un champ d'études avec un fort impact éthique produit le même effet, comme dans le cas des données en danger de disparition des langues minoritaires. Je peux témoigner personnellement d'une valorisation symbolique culturelle à développer un carnet de recherche ouvert à propos précisément de la langue bretonne. Cette valorisation symbolique équilibre dans mon cas, à ce jour, une dévalorisation professionnelle évidente: le temps passé à développer de la science ouverte ne l'est pas à protéger l'accès à ses données mais à inventer des solutions pour accélérer son inverse exact. Ce temps ne me sert pas non plus à créer une liste de publications ouvrant chance à financements de projets. Par-là même, je diminue donc mes chances de pouvoir employer des collaboratrices sur ce même projet, puisque mon profil quantifiable pour les organismes d'évaluation de la recherche reste bas : il est peu clair comment le projet lui-même est évalué (publication de quel rang ? *peer-reviewed* ou non?). Quantifier les citations de l'ouvrage n'est pas non plus une option solide, car l'outil est directement pensé *pour* faciliter sa propre pillabilité, dans un contexte où citer un site web dans un article papier est encore l'exception.

La science à carnets ouverts ne signifie cependant pas en soi l'abandon de toute traçabilité dans la matrice des savoirs. Un site wiki évolue au fil des modifications opérées, mais il est en fait d'une grande traçabilité : il est toujours possible de revenir à une plus ancienne version. Le produit « fini » est un palimpseste qui s'assume, et à chaque page est associé un historique qui retrace toutes les modifications opérée sur cette page depuis sa création. Dans un cas, alors que j'opérais une recension d'article pour un journal, j'ai utilisé cette fonctionnalité du site pour montrer qu'à une date antérieure, une information précise était en ligne sur le site. La pillabilité est donc accrue, mais traçable. La nouveauté du média wiki peut laisser une impression de

recouvrement perpétuel, où la matière contribuée pourrait disparaître, engloutie par des contributions postérieures. Dans la réalité de maniement, toute contribution peut être défaite par l'équipe administratrice, et toute version ancienne peut être aisément retrouvée, et, comme le reste, est consultable en accès libre.

Nielsen (2012) compare l'action de développer une recherche ouverte pionnière au choix de conduire à droite lorsque tout le monde conduit à gauche : penser qu'on peut se remettre à l'initiative personnelle de quelques individus persuadés que ce changement est fondamentalement souhaitable n'est pas un modèle stratégique viable au niveau global. Ces choix doivent être décidés au niveau politique global afin d'assurer une synchronicité nécessaire au projet. Des impulsions politiques ont déjà par le passé été rédhibitoires dans l'établissement de pratiques de science ouverte dans tout un champ. On peut ainsi tracer l'établissement de cartes du génome humain en accès libre à un accord politique impulsé par le milieu scientifique (accord des Bermudes, 1996, Nielsen 2012 :7). Le politique actuellement ne fait pas ces choix. Au niveau européen, les organismes financeurs de la recherche commencent tout juste à introduire, dans certains domaines, un prérequis de financement des projets qui consiste à restituer, en fin de projet, les résultats sous une forme synthétique accessible. Dans l'Etat français en 2013, le politique opère aussi des choix clairs contre l'ouverture des données de recherche. Le discours du politique (discours politiques publics, éditos du journal du CNRS, directives des délégations, stages proposés aux chercheuses, etc.) centre volontairement la valorisation de la recherche sur la privatisation des données via (i) les brevets et (ii) la recherche financée par le domaine privé des entreprises, qui ont rarement intérêt à l'accès libre de ce pour quoi elles doivent payer. Le nouveau modèle de financement de la recherche au coup par coup, par projet, a aussi un effet inhibiteur puissant pour la science ouverte, et ce pour une raison très simple : les personnes les plus à même de les développer sont entretenues volontairement dans une situation de fragilité professionnelle en compétitivité aigue, situation qui incite fortement les compétiteurs à la privatisation des données. Le modèle compétitif prédit que les chercheuses les plus créatives sachant parler à l'institution seront sélectionnés. Ce pourrait donc être une bonne chose, mais ce modèle compétitif prédit aussi que les chercheuses les plus collaboratives seront évacuées du système de recherche. Pour le développement de la science ouverte, c'est un frein considérable.

Ce frein politiquement construit est de plus renforcé par une réalité démographique et culturelle. Le paysage de la recherche est intensément marqué par une fracture numérique générationnelle : les générations de chercheuses avec une sécurité d'emploi et la possibilité de mener des projets collaboratifs à long terme sont les générations les plus éloignées culturellement de la manipulation de l'outil informatique. Les personnels de recherche les plus expérimentés et les plus en position de pouvoir impulser des directions de pratiques de recherche sont précisément les personnes qui ont commencé à faire de la recherche lorsque l'outil informatique était en Europe uniquement anecdotiquement accessible aux particuliers. Les scientifiques qui ont construit la matrice de leurs pratiques de recherche au cœur des techniques informatiques sont donc précisément ceux qui sont sélectionnés par les modèles compétitifs. Je peux témoigner sans équivoque en tant que jeune chercheuse au CNRS que c'est la position

exceptionnelle de sécurité de l'emploi où je me trouve qui me permet de développer un projet de science à carnets ouverts. Si je devais candidater à des postes tous les deux ans, dont la teneur est en partie de rédiger des projets destinés à financer la recherche, une réalisation comme le centre de ressources ARBRES serait probablement réduit à une bibliographie en ligne. Dans le contexte actuel, l'intérêt individuel des chercheuses en contexte de compétition aigue est dans la protection et l'opacité du travail, des données aux hypothèses finales. Le défi à relever est précisément d'inventer des solutions pour aligner les intérêts individuels des jeunes chercheuses avec l'intérêt collectif de développement de la science ouverte. Dans ce défi, les Etats qui ont un service public de recherche pourraient avoir une longueur d'avance.

References

Allen Institute for Brain Science. 2004-2012., *Allen Brain Atlas*, http://www.brain-map.org/

A.P.E.C.S, Association Pour l'Etude et la Conservation des Sélaciens. 2013. *Allo Elasmo*. http://www.asso-apecs.org/-Programme-Allo-Elasmo-.html, accédé le 7 mai 2013.

Dawson, Diane. 2012. Science and Technology Resources on the Internet; Open Science and Crowd Science: Selected Sites and Resources, *Issues in Science and Technology Librarianship* 69, http://www.istl.org/12-spring/internet2.html, Université de Saskatchewan.

Dotlatčil, Jakub & Øystein Nilsen. 2009. The others compared to each other, consequences for the theory of reciprocity, *Proceedings of SALT XVIII*, 2008.

Nielsen, Michael 2012. *Reinventing Discovery: the new era of Networked Science*, Princeton University Press.

Baider, Fabienne, Edwige Khaznadar et Thérèse Moreau, 2007. Les enjeux de la parité linguistique, *Nouvelles Questions Féministes*, 26 :3, 4-13.

Farchy, Joëlle, Pascal Froissart & Cécile Méadel. 2010. Science.com, libre accès et science ouverte, *Hermès* 57, Paris : CNRS Éditions.

Howe, J. 2006. The rise of crowdsourcing, *Wired* 14.06**.,** http://www.wired.com/wired/archive/14.06/crowds.html, accédé le 6 mai 2013.

InnoCentive. 2013. http://www.innocentive.com/, accédé le 7 mai 2013.

Jackson, Wendy & Joel Brown. 1997-2011. *Project Squirrel*. http://www.projectsquirrel.org/lessons.shtml, Université de Chicago et Académie des sciences de Chicago.

Jouitteau, Mélanie. 2009-2013. *ARBRES, Vers un Atlas de la microvariation dialectale de la langue bretonne*, http://arbres.iker.univ-pau.fr/index.php/Accueil.

Khaznadar, Edwige. 2007. Le non-genre académique: doctrine de la domination masculine en France, *Nouvelles Questions Féministes*, 26 :3, 25-38.

Lossec, Herve. 2010. *Les bretonnismes*, Skol Vreizh.

Nielsen, Michael. 2012. *Reinventing discovery*, Princeton University Press.

Ocean Observatories Initiative: http://www.oceanobservatories.org/, accédé le 7 mai 2013.

Popović, Zoran & David Baker. 2007-2013. *Foldit, Solve puzzles for science*. http://fold.it/portal/

Project BudBurst. 2013. *Project BudBurst: An online database of plant phenological observations*. Project BudBurst, Boulder, Colorado. Available: http://www.budburst.org; Community Attribution: http://budburst.org/attribution.php; Accessed: April 26, 2013.

Van der Merwe, M., J.S. Brown et W.M. Jackson. 2005. The Coexistence of Fox (Sciurus niger) and Gray (S. caroliniensis) Squirrels in the Chicago Metropolitan Area, *Urban Ecosystems* 8:335-347.

Salganik, Matthew J. & Karen E. C. Levy. 2012. *Wiki surveys: Open and quantifiable social data collection*. manuscrit : http://arxiv.org/abs/1202.0500

Schöpfel, Joachim. 2011. Science.com, libre accès et science ouverte, *Bulletin des Bibliothèques de France* 56 : 5. 122-123. [en ligne] <http://bbf.enssib.fr/> Consulté le 28 avril 2013.

Starke, Michal. 2004-2013. *Lingbuzz*, http://ling.auf.net/lingbuzz.

Surowiecki, James. 2008. *La Sagesse des foules*, éditions Jean-Claude Lattès.

Surowiecki, James. 2004. *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economics, Societies and Nations*. New York: Doubleday.

[i] Cet article est écrit au féminin général, mais il est bien entendu que les lecteurs et contributeurs du site sont concernés de façon tout à fait identique. La raison en est un souci simple d'ouverture des outils nouveaux aux populations les plus concernées par la fracture numérique, en cohérence avec l'essence même du projet ARBRES. Pour une discussion éclairée de ces enjeux, se reporter à Baider, Khaznadar et Moreau (2007) et Khaznadar (2007). Le lecteur ou la lectrice peut se faire une idée approximative de la fracture genrée en sciences ouvertes au travers de la bibliographie et webliographie du présent article, ou des listes d'équipes de direction des projets mentionnés.
Merci à Juan Baztan, Michal Starke, Manon Jouitteau, Béatrice Bazin, Loïc Jade, Lorent Moineau et Henriette Brossas pour leurs commentaires sur une version antérieure de ce texte. Je reste responsable de toute erreur ou approximation restante.

[ii] La langue utilisée est aussi un facteur aggravant: les publications les plus valorisées dans les publications linguistiques sont de langue anglaise, langue qui est peu accessible en Bretagne, spécifiquement dans les tranches d'âge bilingues en français/breton. La science ouverte ne résout pas ce problème d'accessibilité. ARBRES est rédigé en français pour cette raison.

[iii] J'ai personnellement poursuivi le projet avec Michal Starke de créer un site frère de *Lingbuzz*, plateforme qui distribuerait les archives multilingues des publications concernant la langue bretonne. J'ai dû abandonner ce projet faute de connaissances informatiques assez précises pour entrer dans les lignes de code et adapter le programme. Dans le sein de l'Etat français, un tel projet a été reçu comme potentiellement concurrentiel à HAL et n'a pu voir le jour. Il n'existe pas à ce jour de lieu sur le web où des personnes intéressées par la langue trouveraient accès aux dernières publications.

[iv] SSWL passera en 2013 sur http://www.terraling.com/, avec un remaniement de l'interface.