Des insouciances de l'archivisme à une réflexivité constante : considérations éthiques en linguistique de terrain à l'ère du Traitement Automatique des Langues

Camille Noûs, Alexis Michaud

Des insouciances de l'archivisme à une réflexivité constante : considérations éthiques en linguistique de terrain à l'ère du Traitement Automatique des Langues

Camille Noûs

Laboratoire Cogitamus

Camille Noûs est un individu collectif qui symbolise notre attachement aux valeurs d'éthique et de probation que porte le débat contradictoire, et conscient de ce que nos résultats doivent à la construction collective ; le « Noûs », porteur d'un Nous collégial, est également une référence au concept d'« intellect » (νοῦς, *noûs*) de la philosophie grecque.

Alexis Michaud

Langues et Civilisations à Tradition Orale (LACITO), CNRS – Université Sorbonne Nouvelle, Villejuif, France

Linguiste de terrain, passionné par la documentation des langues rares et menacées et par les questions liées à l'archivage pérenne et la diffusion des corpus.

identifiant ORCID : 0000-0003-1165-2680

alexis.michaud@cnrs.fr

Ce texte se veut un essai de retour réflexif sur un parcours : deux décennies d'enquêtes linguistiques de terrain sur des langues rares, au Yunnan (Chine), et d'engagement en faveur de l'ouverture des données de terrain (corpus multimédia de langues rares). Le cadre temporel coïncide avec deux décennies de progrès spectaculaires de l'informatique, et notamment du Traitement Automatique des Langues. Au fil des réflexions, il ressort que l'adoption conséquente de principes de Science ouverte amène, non pas à se doter d'un ensemble de solutions qui permettent d'éviter les soucis, mais à faire une place centrale aux questions éthiques et socio-politiques soulevées par la constitution, la publication électronique et l'exploitation de ressources en langues rares. Ces thèmes appellent une réflexion collective constante, à laquelle soient associé·es des spécialistes de Traitement Automatique des Langues.

Mots-clés : données ouvertes, langues rares, archives, éthique et déontologie, Traitement Automatique des Langues

From Archival Activism to Constant Reflexivity: Ethical Considerations in Field Linguistics in the Era of Natural Language Processing

This text is an attempt to look back on two decades of linguistic fieldwork on endangered languages in Yunnan (China) and of activism in favour of opening up fieldwork data (multimedia corpora of undocumented or little-documented languages). The time frame coincides with two decades of spectacular progress in computer science, and in particular in Natural Language Processing. Looking back, it seems clear that the consistent adoption of Open Science principles implies that the ethical and socio-political issues raised by the creation, electronic publication and exploitation of fieldwork data must be given a central place (rather than hoping for ready-made solutions to deflect concerns and fend off trouble). These topics call for constant collective reflection involving specialists in Natural Language Processing.

Keywords: open data, linguistic heritage and diversity, archives, research ethics, Natural Language Processing

从无所忌惮的数据存档激进主义到不断的反思：自动语言处理时代田野语言学的伦理考量

本文试图回顾二十年来在云南（中国）开展的濒危语言田野调查，以及为开放田野调查数据（未记录或记录较少的语言多媒体语料库）所做的努力。这二十年，是计算机科学，特别是自然语言处理技术取得巨大进步的二十年。回顾过去，我们似乎可以清楚地看到，始终如一地采用开放科学原则意味着：必须将实地调查数据的创建、电子出版和利用所引发的伦理和社会政治问题放在中心位置。而不是寄希望于依靠现成的解决方案来转移人们的关注和避免纠纷。这些问题的解决，需要语言学家与语言处理专家们协同合作，不断进行集体思考。

Introduction

Le présent texte[1] se veut un essai de retour réflexif sur le volet juridique et éthique d'un parcours de linguiste de terrain : deux décennies d'enquêtes sur des langues minoritaires (en voie de remplacement par la langue nationale) au Yunnan (Chine), de publication de ressources orales en archive ouverte, et d'exploitation de ces ressources. « Les politiques de science ouverte ont placé au cœur du travail scientifique l'enjeu de l'ouverture des données de la recherche (Galonnier *et al.* 2019). (…) Cependant, très peu de publications rendent compte de manière documentée de l'expérience de chercheurs ou d'ingénieurs, pourtant les premiers concernés (…) [L]'intérêt à restituer certaines expériences pour réfléchir collectivement à ces questions est réel et invite à décaler la perspective par rapport au seul cadre théorique des « bonnes pratiques » » (Bendjaballah et Garcia 2023). Au fil des réflexions, il ressort que le militantisme en faveur de l'ouverture des données de la recherche ne constitue pas par lui-même une réponse complète aux questions soulevées par la constitution, la publication électronique et l'exploitation de ressources linguistiques en langues rares. Sur la base de ce constat (largement partagé) ressortent quelques pistes qui constituent autant de chantiers pour les Communs numériques.

Les insouciances de l'archivisme numérique à la naissance d'internet : l'ouverture des données comme mouvement à sens unique

Les années 1990 ont été marquées par une double révolution pour la conservation et la diffusion de données de langues rares. D'une part, la numérisation rendait théoriquement possible la conservation indéfinie des données multimédia, puisque la copie numérique se fait à l'identique, alors que les supports analogiques vieillissent de façon lente mais irréversible (Bonnemason, Ginouvès et Pérennou 2001, 4). D'autre part, ces ressources pouvaient être partagées avec le monde entier, grâce à internet. D'Europe de l'Ouest, on pouvait aller partout (la réciproque était moins vraie) ; la Corée du Nord et quelques zones de conflit figuraient des îlots anachroniques, fermés aux voyageurs et à internet, dans un monde ouvert. La mondialisation laminait les sociétés, les cultures et les langues, mais il y avait quelque chose à faire pour sauver du naufrage un peu du patrimoine linguistique mondial.

Travailler à une archive numérique avant l'essor de la Science ouverte : le temps des pionniers

Jusqu'aux années 2010, notre engagement pour l'archivage numérique (« archivisme numérique ») était marqué par une relative insouciance[2] vis-à-vis des questions éthiques. Il s'agissait d'abord de sauver les données, en bâtissant une archive ouverte numérique afin de mettre fin à une situation de déperdition quasi-généralisée des données de la recherche.

> Les chercheurs et étudiants ont tendance à constituer leur propre corpus à mesure des besoins de leur recherche, plutôt que de raisonner en termes de patrimoine documentaire partagé. Les fonds d'archives sont peu connus, les grands corpus distribués sur internet dépassent souvent les budgets du chercheur individuel, tandis que l'on peut enregistrer soi-même un corpus d'une qualité technique satisfaisante.

1 Le choix de faire figurer Camille Noûs comme autrice symbolique de cet article (suivant en cela l'exemple d'une collègue de mon laboratoire : Simon et Noûs 2021) vise à souligner la nature collégiale du travail scientifique. Camille Noûs est un individu collectif qui symbolise notre attachement aux valeurs d'éthique et de probation que porte le débat contradictoire.

2 Le terme d'*insouciance* est une allusion à une communication dans laquelle le linguiste Michel Ferlus revient sur le parcours qui a mené de la collecte de données jusqu'à leur publication numérique, non prévue initialement : « Des enregistrements sans expérience à la numérisation : les insouciances du terrain » (Ferlus 2017).

> On voudrait souligner ici les limites de cette logique : il est illusoire de penser que l'on peut à tout moment créer le corpus dont on a besoin. Dans le cas des langues en danger, la mise en commun des données existantes est particulièrement nécessaire. Mais dans l'étude des grandes langues, le travail documentaire ne demande pas moins de sérieux. (Michaud 2002)

Comme on l'imagine, pareil cours de morale dispensé par un nouveau venu drapé dans des habits de bonne sœur à cornettes ne recueillait guère l'adhésion des collègues. « Les chercheurs étaient recrutés pour faire des recherches, publier des résultats mais, pour la majorité d'entre eux, ils ne se considéraient pas investis d'une mission de service public pour la diffusion des données qu'ils collectaient » (Simonnot 2020). Le programme Archivage du LACITO (Jacobson, Michailovsky et Lowe 2001), qui prendrait plus tard le nom de « collection Pangloss », était opérationnel dès la fin des années 1990, pour pérenniser, exploiter et diffuser les enregistrements des linguistes de terrain. Annotation balisée selon la norme XML et synchronisée phrase par phrase avec l'enregistrement numérisé, consultation au moyen d'un simple navigateur internet : tout paraissait en place pour emporter l'adhésion des linguistes. Mais à notre grande surprise, l'effet d'entraînement tardait (bien que le site ait attiré l'attention de quelques collègues[3]). Le travail de mise en forme et archivage de données était activement découragé par des collègues qui jugeaient que ces tâches de technicien-ingénieur et non de chercheur ralentissaient d'autant le rythme de publication, qui seul comptait dans l'évaluation.

C'était le règne du « mérite publicationnel » jugé à l'aune de la bibliométrie à l'américaine (statistique des citations). Face au constat des limites de cet outil (Archambault et Larivière 2009), des propositions de rafistolages divers et variés se succédaient (Fassoulaki *et al.* 2002), sans remettre en cause le postulat selon lequel la solution, la bonne, serait une mesure quantitative enfin débiaisée. Pour l'heure, l'institution retirait son soutien financier aux revues de « catégorie C » (Gunthert 2004) et s'autorisait à modifier le classement proposé par les instances de recrutement lorsqu'il ne coïncidait pas avec les indicateurs quantitatifs.

Dans un tel contexte, les occasions de se frotter aux questions juridiques et éthiques liées aux données étaient rares, et ces thèmes n'étaient guère abordés dans les cursus universitaires. Le mouvement pour la constitution d'archives ouvertes se concentrait sur la mise en réseau d'archives numériques au plan mondial ; les thèmes juridiques et éthiques ne figuraient pas parmi les questions stratégiques identifiées (Simons et Bird 2000). Bref, l'ouverture des données était un « mouvement à sens unique », pour reprendre l'expression que Frédérique Bordignon applique au mouvement en faveur du libre accès aux publications (Bordignon 2023).

Une certaine défiance vis-à-vis de l'approche nord-américaine des questions d'éthique de la recherche

Le cadre juridique indispensable à l'activité de publication de données était en place dans l'archive ouverte, mais n'apparaissait aux chercheurs que de façon discrète, voire marginale. À l'étape du dépôt, le chercheur déposant apprenait être identifié comme détenteur du droit de copie *(copyright)* ; choix valorisant et rassurant, en cohérence avec le rôle central qu'accorde le CNRS au chercheur, « tête pensante » autour de laquelle s'organise l'institution, et qui bénéficie des services assurés par les personnels d'*appui à la recherche*. On se voyait proposer un choix quant à la licence sous laquelle étaient placés les documents déposés. (La collection Pangloss et la plate-forme Cocoon qui l'héberge ont très tôt recommandé les licences Creative Commons.) Il n'était pas nécessaire de fournir de document attestant du consentement des personnes enregistrées. Cela revenait à faire confiance au déposant pour les questions juridiques et éthiques liées à la collecte, l'archivage et la diffusion de données de terrain.

3 Notamment Nick Thieberger, qui posait à l'époque en Australie les bases de ce qui est devenu l'une des principales archives de langues rares : *Pacific and Regional Archive for Digital Sources in Endangered Cultures* (PARADISEC) (voir Harris, Thieberger et Barwick 2015 ; et, en français, Vernaudon *et al.* 2021, 327).

Les pratiques sur le terrain étaient teintées d'une nette défiance vis-à-vis des procédures de validation éthique qui avaient cours dans le contexte nord-américain. On pressentait bien que l'Europe finirait par suivre le mouvement, mais on renâclait à l'idée de devoir emboîter le pas à une Amérique qui, en matière de morale, ne manifestait pas un grand sens de la nuance.

> La Russie n'apportait plus d'espoir ni de peur, rien d'autre qu'une désolation perpétuelle. Elle s'était retirée de notre imaginaire – que les Américains occupaient malgré nous, comme un arbre gigantesque étalant ses branches à la surface de la terre. Ils nous énervaient de plus en plus avec leur discours moral, leurs actionnaires et leurs fonds de pension, leur pollution de la planète et leur dégoût de nos fromages. (…) Des conquérants sans idéal sinon le pétrole et les dollars. Leurs valeurs et leurs principes – ne compter que sur soi – ne donnaient d'espérance à personne d'autre qu'eux et nous rêvions d'« un autre monde ». (Annie Ernaut, *Les Années*, 2008, p. 209)

Les « autres mondes » culturels, la linguistique de terrain a précisément pour objet d'aller à leur rencontre, et de prendre la mesure de leur diversité socio-politique, aisément méconnue (Woodbury 2011). Or une réglementation en matière d'éthique de la recherche ne peut qu'être en décalage plus ou moins profond avec les situations concrètes du terrain. Les réactions de locutrices et locuteurs prié·es de signer des formulaires de consentement éclairé approuvés par un Comité d'éthique de la recherche (les exemples que j'ai en tête, employés au Vietnam, provenaient d'Allemagne et des États-Unis) me paraissaient éloquentes : le document (parfois une liasse comportant plus de mille mots) n'était pas lu, tout au plus rapidement survolé. La signature était précédée d'une brève explication au sujet de ce que représentait le document et de son caractère indispensable. La personne appelée à signer réagissait parfois par un simple regard de connivence *(allez on est sympa, on te le signe)*, parfois en relevant le caractère vaguement inquiétant de la procédure *(C'est bien alambiqué, tout ça ! Bon, tiens, voilà)*. L'étape du consentement constituait une formalité, perçue comme telle. Cela faisait penser à du *consent washing* (ou *ethics washing*), comme le *greenwashing* qui donne une image de responsabilité écologique pour pas trop cher.

Dans le contexte de cultures à tradition orale, il pouvait paraître plus respectueux d'éviter cette formalité aux personnes qui nous enseignent leur langue et nous autorisent à recueillir leur parole. Dans les localités du Yunnan où s'est déroulé mon terrain, les enfants sont scolarisés en chinois, et chaque citoyen·ne est amené·e à signer des documents à certaines étapes de la vie. Mais les écrits légaux dont est familière la culture locale engagent des décisions importantes, de sorte que l'acte de signer comporte des connotations culturelles appartenant à un registre bien éloigné de celui qui est souhaitable pour une enquête linguistique. Là-bas comme ici, la sagesse populaire enseigne qu'il est bien préférable de n'avoir jamais affaire à la justice.

On peut aller jusqu'à reprocher aux protocoles éthiques des tendances impérialistes, comme le relève un spécialiste de bioéthique. « I think the amateurs in ethics suffer in imperialism. They practice imperialism, which is constituted by establishing a rule and applying it ruthlessly. You know, wherever you go, you set your rule down. What we ought to be particularly good at by now is seeing that rules have very different meanings as situations approach or depart from a kind of a paradigm situation », observe Albert Jonsen, cité par Zachary Schrag (2010, 7). De même que, selon la formule pascalienne, *la vraie morale se moque de la morale*, l'éthique de la science se moquerait de l'éthique formaliste, dont la fonction première serait tout bonnement – loin des préoccupations déontologiques affichées – de protéger l'institution contre le risque de coûteux procès (Bowern 2010, 900 ; van Driem 2016). La vraie responsabilité éthique consiste à entretenir les meilleures relations avec les consultants linguistiques : adopter un comportement bien adapté au contexte culturel, respecter les consultants, valoriser les connaissances qu'ils partagent, leur expliquer l'ensemble du processus de documentation et de recherche, leur donner une juste compensation pour le temps qu'ils consacrent au travail, leur permettre d'accéder à la documentation linguistique produite ainsi qu'aux productions du chercheur. Et les associer au processus de documentation et de recherche, avec l'espoir que le consultant linguistique devienne un collaborateur et, idéalement, un collègue. Tout cela est exposé dans un ouvrage de référence

(Bouquiaux et Thomas 1971) dont on m'a remis un exemplaire lorsque j'ai rejoint le laboratoire de Langues et civilisations à tradition orale (LACITO, fondé en 1976), et que j'ai, depuis, recommandé à mon tour à des doctorantes et doctorants.

En outre, l'étude de langues minoritaires dans des pays dont les politiques assimilationnistes (Gros 2014) excluent la mise en place de filières de scolarisation en langue maternelle (pourtant très positives au plan éducatif), et visent à la généralisation de la langue nationale, ne pouvait guère prendre le tournant collaboratif et communautaire dont des projets nord-américains ou australiens donnent l'exemple (Snead et Cushman 2023).

> Depuis deux ou trois décennies, les préoccupations éthiques prennent une importance croissante dans le travail des linguistes. Il est en effet difficile de ne considérer que sous l'angle de l'intérêt intellectuel l'étude d'une langue en danger de disparition, ou au moins socialement dévalorisée, et dont les locuteurs sont parfois déplacés de leur lieu et de leur mode de vie traditionnels, fragilisés par la rupture avec leurs traditions culturelles, et relégués dans les marges défavorisées de la société. Des enquêtes internationales – malheureusement peu relayées en France – menées sur des populations autochtones d'Amérique, d'Australie ou des pays arctiques, montrent que la perte des langues ancestrales est l'un des éléments d'un mal-être général, qui se traduit aussi par des problèmes de santé (diabète, obésité, maladies cardiaques) et des comportements à risques comme l'alcoolisme, l'addiction aux drogues, le suicide, la délinquance, alors que le maintien des langues et des références culturelles réduit considérablement ces dangers. Il est difficile pour les linguistes de rester indifférents à ce constat, ou même de s'en tenir à une empathie compassionnelle, sans s'interroger sur la place qu'ils occupent dans un tel contexte, et sur les moyens d'y jouer un rôle positif. (Launey 2023, 730-731)

La relative discrétion dans laquelle se déroulaient mes premières enquêtes linguistiques au Yunnan, dans le sud-ouest de la Chine, dans les années 2000, limitait les échanges au sujet du projet documentaire et scientifique avec des membres de la communauté linguistique concernée. Or le fait que des locutrices et locuteurs soient partie prenante au projet de documentation linguistique figure parmi les principaux moteurs de réflexion sur les questions d'éthique, au sens large (Vapnarsky 2020 ; Marsault 2021, 27-38 ; Cox 2022). Le choix du nom donné à une langue (Bichurina 2017, 500-501 ; Michaud 2017c, 496-501), les décisions prises au stade crucial que constitue la mise à l'écrit d'une langue à tradition orale, comportent toutes sortes d'enjeux de politique linguistique, qui engagent la définition même de la communauté et soulèvent la question de la légitimité de telles ou telles personnes, en tant que membres, en tant que représentants, en tant que chercheurs « accrédités »… Ces thématiques, délicates à gérer (Launey 2023, 731-732), mettent les questions éthiques sur le devant de la scène, amenant à des prises de position circonstanciées de la part des praticiennes et praticiens de la linguistique de terrain et de l'étude des langues rares. Ainsi, l'autrice et l'auteur d'une édition trilingue inuktitut-anglais-français de l'ouvrage *Chasseur au harpon*, de l'écrivain inuk Markoosie Patsauq, ont été amenés à mener une réflexion critique au sujet des risques de l'entreprise (essentialisation de la langue, perpétuation de schémas coloniaux) et des moyens à déployer pour les déjouer, à commencer par des échanges suivis et équilibrés avec les membres des communautés inuits (Henitiuk et Mahieu 2024, 3-4).

Rétrospectivement, il ressort que ma pratique d'une ouverture des données « à sens unique » (plutôt que dans un processus collectif, réflexif et continu) n'était pas sans lien avec les parois de verre qui maintenaient une certaine distance entre la communauté des locutrices et locuteurs – dont le représentant légitime, dans un État à parti unique, n'est autre que ce parti – et le processus de constitution et exploitation des ressources produites. Sur le terrain, je ne faisais pas signer de document par les locutrices et locuteurs. Lorsque je me suis finalement essayé à recueillir un consentement sur le terrain, l'essai ne m'a pas paru concluant.

Autorisation de diffusion et consentement oral : un essai peu concluant

En dépit des réticences exposées ci-dessus, j'étais désireux de disposer de tous les arguments possibles en faveur du libre accès que je souhaitais pour les données que je recueillais. J'éprouvais la crainte qu'en l'absence de preuve de consentement, il ne me soit, à plus ou moins long terme, reproché d'avoir manqué aux « bonnes pratiques ». De fait, à la date de rédaction du présent texte (2024), la charge de la preuve incombe au chercheur, qui doit disposer d'éléments établissant (par exemple auprès d'une revue à laquelle on soumet un article) que la collecte de données répond aux exigences éthiques en vigueur. Un adage romain veut, paraît-il, qu'en matière de droit, *idem est non esse aut non probari* : ne pas être, ou ne pas être prouvé, c'est tout un.

J'ai donc entrepris, au début des années 2010, d'éliciter un consentement, par écrit et par oral, de la personne qui m'enseignait la langue na de Yongning depuis 2006. Pour le document écrit (en chinois, le na étant une langue à tradition orale), un membre de la famille a signé pour la locutrice, pis-aller qui souligne la fragilité de l'exercice, et qui cause de l'embarras à l'intéressée, en soulignant qu'elle n'a pas été « écolée »[4] – alors qu'il importerait au contraire de valoriser les précieux savoirs dont la personne est détentrice, et qu'elle accepte de partager.

L'enregistrement d'un consentement oral permet de replacer le propos dans l'espace de l'oralité et de la langue étudiée, mais la mise en œuvre révèle d'autres fossés culturels. Mon idée était d'enregistrer un court message par lequel la locutrice marquerait son accord pour que n'importe qui puisse écouter les enregistrements et en savoir plus sur sa langue. Pour cet exercice imposé, il lui était demandé de s'adresser, dans sa propre langue, à un public abstrait qui s'intéresserait au processus d'enquête et à la relation de travail que nous avions nouée. Ma prof a accédé à cette drôle de demande, qui lui avait été expliquée par moi-même et par son fils. Ce qu'elle a finalement dit[5] est, en substance : « Voilà maintenant trois, quatre ans que Diddeo [le nom qui m'a été donné en langue na] est venu chez moi. C'est mon fils qui le connaissait ; mon fils m'a présenté Diddeo pour que je lui enseigne la langue na ; je lui ai dit tout ce que je savais ; est-ce qu'il est calé ou pas, ça, c'est vous qui savez. Depuis trois ou quatre ans qu'il étudie, il comprend un peu la langue ; moi je parle, lui il écrit ; il est consciencieux, c'est quelqu'un de calé. Avec les enregistrements, vous êtes en mesure d'écouter, et de voir si vous arrivez à comprendre tout ça ». L'enregistrement, qui comporte par endroits une discrète note d'humour, pourrait être considéré comme trop éloigné du contenu attendu pour remplir la fonction souhaitée : établir la conformité de l'enquête linguistique avec la réglementation en matière d'éthique. Cet essai de recueil de consentement ne paraissait pas concluant : l'éthique réglementaire ne semblait décidément pas faire bon ménage, en pratique, avec l'éthique de la recherche.

Numérisation de sauvegarde de fonds anciens : priorité à l'ouverture

Dans le travail de sauvegarde de données anciennes (des fonds irremplaçables, qui dans bien des cas disparaissent avec leur auteur), la priorité allait pareillement à l'ouverture des données, plutôt qu'à l'approfondissement des questions juridiques et éthiques soulevées par l'entrée des données dans le monde numérique. Ainsi de la numérisation des collections d'enregistrements audio de Michel Ferlus, spécialiste des langues d'Asie orientale, réalisée avec un financement du Comité pour la Science ouverte (anciennement Bibliothèque scientifique numérique) (Michaud 2017b). Dans le cadre de l'Appel à projets « Numérisation du patrimoine », un des critères d'admissibilité était l'engagement à mettre en ligne les documents numérisés, en libre accès, au plus tard à la date de fin du projet. Cette exigence a contribué à convaincre Michel Ferlus, initialement réticent à l'idée d'ouvrir au public des enregistrements qui à l'origine étaient réalisés comme simples documents de travail (Ferlus, 2017), de donner son accord à la diffusion sous licence Creative

4 J'emprunte ce terme désuet à une grand-mère tenue à l'écart de l'école, et qui mesurait l'importance que les femmes fussent « écolées ».

5 Le document est disponible ici : https://doi.org/10.24397/pangloss-0004611. Chacune des ressources de la collection Pangloss bénéficie d'un identifiant DOI : voir Vasile et al. (2020).

Commons (BY-NC-SA)[6]. La question du consentement des personnes concernées paraissait anachronique, s'agissant de données collectées depuis les années 1960, sur plus de trente langues et dialectes de la péninsule indochinoise (Laos, Vietnam, Cambodge, Birmanie), généralement dans des localités difficiles d'accès. Recueillir un consentement *a posteriori* paraissait infaisable, et au fond assez futile en comparaison de l'enjeu de la conservation et l'ouverture des données.

Un changement de perspective : le droit et l'éthique comme alliés pour l'ouverture des données de la recherche en sciences humaines et sociales[7]

À partir du milieu des années 2010, j'ai eu la chance d'avoir des occasions d'échanges avec Danièle Bourcier, introductrice en France des licences Creative Commons, puis, grâce à elle, avec Lionel Maurel, juriste de formation, conservateur des bibliothèques, qui s'intéresse comme elle aux biens communs, à la culture libre, au domaine public. Il a fallu ces échanges pour que je prenne conscience des travers que comporte l'attitude paternaliste qui consiste à se poser en protecteur d'une communauté fragile (indigènes, minorité linguistique, groupe social marginalisé…). À se raidir contre le principe d'une formalisation du consentement, on a vite faite de s'arroger sur les gens une autorité morale, de leur imaginer un devoir de gratitude proportionnel aux bontés qu'on estime avoir pour eux, et de se bâtir un fief scientifique (chasse gardée). Une autre dérive consiste à invoquer le principe de protection de la vie privée pour légitimer le fait de garder par-devers soi les enregistrements, qui seront, au final, tout bonnement perdus. Certes, les enregistrements sont recueillis dans le cadre d'une relation de confiance personnelle avec l'enquêteur[8]. Pour autant, il n'est pas du tout évident que leurs auteurs souhaitent limiter à cette personne la circulation d'enregistrements qui ont pour objet les progrès de la connaissance.

Si incongru que cela puisse paraître *a priori*, il n'y a en réalité rien d'inconvenant à avoir, pendant une enquête linguistique de terrain, des conversations au sujet de thèmes juridiques tels que les licences sous lesquelles on prévoit de placer des enregistrements. En expliquant la différence entre une licence Creative Commons et un contrat de cession de droits patrimoniaux à titre exclusif, on contribue au partage d'une information juridique qui n'est pas sans valeur pour les citoyens concernés, et on facilite un débat au sujet de la question du partage des données.

Les réflexions ainsi ouvertes se poursuivent depuis, dans un contexte institutionnel où les pratiques de Science ouverte sont désormais encouragées et soutenues.

De 2016 à nos jours : vers une généralisation des pratiques de Science ouverte

Dans la seconde moitié de la décennie 2010, la Loi pour une République numérique changeait la donne en matière de données de la recherche : l'ouverture des données ne serait plus un choix personnel militant, mais un devoir professionnel partagé (Arènes, Maurel et Rennes 2022). C'était là un bouleversement considérable : la carotte et le bâton allaient, bien mieux que les arguments de principe au sujet des vertus de la Science ouverte, convaincre enfin l'ensemble des chercheuses et chercheurs d'élaborer un plan de gestion des données, et de prendre le chemin de la publication

6 Cette observation fait écho à celle effectuée par Joséphine Simonnot au sujet des numérisations effectuées avec le soutien de la Mission Recherche et Technologie (MRT), qui avait fixé pour politique que les financements de la numérisation devaient s'inscrire dans le cadre de projets de diffusion au public (Simonnot 2020, §3).

7 Ce titre reprend l'intitulé d'une journée d'étude organisée en 2019 par le groupe de travail Éthique & Droit du Comité pour la Science ouverte à la Maison méditerranéenne des sciences de l'homme (MMSH) d'Aix-en-Provence : « Diffuser les données numériques en SHS : le droit et l'éthique comme alliés ».

8 La personne qui m'a enseigné la langue na de Yongning est présentée dans le livre que j'ai écrit au sujet du système tonal de la langue (Michaud 2017c, 29-33). De tristes circonstances m'ont également amené à écrire un mot au sujet de ma première consultante de langue naxi (Michaud 2017a).

électronique des données de terrain. Documentation et recherche allaient enfin progresser de façon solidaire. Une équipe de douze autrices et auteurs se constituait pour recommander une transition vers une politique de données ouvertes dans les sciences phonétiques.

> …the phonetic sciences stand to gain greatly from data availability and citability (…) Our hope is to facilitate a base level of common understanding so that the field can deal with these core issues actively and manage ongoing transitions tactfully, rather than passively letting changes happen around us and belatedly realizing that data have evaporated and many research articles in phonetics play to our ears "ditties of no tone", like the silent Grecian urn in John Keats's *Ode* (Garellek *et al.* 2020, 3-4).

L'historique éditorial de cette prise de position collective, publiée dans une jeune revue brésilienne (en libre accès diamant) après avoir été rejetée par une revue-phare du domaine, rappelait que presque rien n'avait bougé en deux décennies. La transition ne se ferait pas du jour au lendemain, mais c'était une raison de plus d'y aller de bon cœur. Il y avait du pain sur la planche, et il gardait une saveur de militantisme.

Inscription au registre des bases de données du CNRS : échanges avec la Déléguée à la Protection des Données

L'inscription de la collection Pangloss au registre des bases de données du CNRS, préparée avec Gaëlle Bujan, Déléguée à la Protection des Données du CNRS, a fait ressortir des éléments rassurants. La licéité du traitement de données à caractère personnel se fonde sur le fait que ce traitement est mis en œuvre dans le cadre de l'exécution d'une mission de service public. La sensibilité des données est jugée *Normale* (les deux autres catégories possibles étant *Sensible* et *À risque*), et il n'y a pas d'obstacle à la publication du nom des personnes ayant contribué à la création des ressources (à commencer par les locutrices et locuteurs – lorsqu'il n'y a pas de contre-indication, cela va de soi : les métadonnées sont anonymisées si le contexte socio-politique le demande). Il n'y a donc pas d'obligation d'anonymisation des personnes, qui entrerait en conflit avec la politique de la collection de reconnaître le rôle de chacun (à commencer par les locutrices et locuteurs) dans la constitution des ressources orales[9]. Ces éclaircissements, qui établissent la conformité de la collection Pangloss à la législation, constituent en outre d'utiles éléments de réponse à des questions qui se posent régulièrement dans la formulation de projets de recherche sur des langues rares, projets qui doivent dorénavant être soumis à l'examen d'un Comité d'éthique de la recherche.

Le séminaire « Science ouverte », lieu de formation et d'échange

Un séminaire « Science ouverte »[10] organisé par un collectif du campus CNRS de Villejuif en 2020-2021 et 2021-2022 visait à faire entrer de plain-pied les doctorantes et doctorants (et toutes autres personnes intéressées) dans les pratiques de Science ouverte, retour aux sources de l'*ethos* de la science[11]. L'Université Sorbonne Nouvelle, informée du projet, proposait qu'il soit mutualisé par l'ensemble de ses Écoles doctorales ; le séminaire était également accrédité par l'Inalco. Mastérants, doctorants, chercheurs, ingénieurs, collègues du service juridique de la Délégation CNRS se retrouvaient pour des séances dont la seconde moitié était consacrée à une discussion libre, succédant à un exposé donné par des praticien·nes expérimenté·es. L'argumentaire du

9 Cette perspective rejoint celle de l'ethnographie contemporaine : « [l']évolution des cadres éthiques du métier d'ethnographe de ces dernières décennies (Caplan 2003 ; Fluehr-Lobban 2013) assigne en effet aux communautés le statut de copropriétaire des données résultant de la relation ethnographique, ce qui ouvre la voie à une coresponsabilité de leur préservation et de leur partage entre chercheur et communauté étudiée – et les incite à trouver les moyens de son exercice » (Heintz 2023).

10 Voir : https://himalco.hypotheses.org/

11 L'expression est empruntée à Calimaq : https://scinfolex.com/2019/06/05/louverture-des-donnees-de-recherche-un-retour-aux-sources-de-lethos-de-la-science/

séminaire, calquant une formule de Rousseau, annonçait la résolution d'*allier toujours ce que la Science ouverte recommande avec ce que l'intérêt prescrit, afin que bonnes pratiques et perspectives de développement de carrière ne se trouvent pas divisées*[12]. Est-ce que les pratiques de Science ouverte constitueraient pour les doctorantes et doctorants un atout professionnel, ou est-ce qu'elles leur ouvriraient seulement la petite porte d'emplois précaires et ancillaires, comme gestionnaires des données d'autrui ? Les avis étaient notablement différents d'une discipline à une autre. On convenait de laisser les intéressés faire leurs choix librement, au vu de l'information dispensée, et d'un contexte mondial qui se dégradait à vue d'œil. L'esprit critique universitaire était en butte à des attaques médiatiques (Eslen-Ziya, Giorgi et Ahi 2023) que légitimaient les déclarations de ministres de la République en exercice. On pointait, à notre manière (Michaud, Nguyễn et He 2020), l'erreur d'un président de la République qui disait tout haut ce qu'il pensait des « gens qui ne sont rien ». *Médiapart* résumait l'état des services publics, malmenés par des politiques d'austérité sans fin, par la formule « après l'alerte, la dégringolade ».

À la différence de l'engagement écologico-social, lui aussi fondé sur un constat scientifique clair (Ripple *et al.* 2017), mais qui s'inscrivait sur fond de désespoir (Kaufmann *et al.* 2019) et de répression politique méthodique (Middeldorp et Le Billon 2020), le militantisme pour la Science ouverte avait le soutien de l'institution. Les sciences du langage figuraient à l'avant-garde, fortes de revues et maisons d'édition en libre accès diamant (sans frais pour les autrices ni les lectrices), comme *Glossa* (Rooryck 2016) et Language Science Press[13], et de plate-formes spécialisées pour les données, telles que Cocoon[14]. On était invité à raconter sur un blog institutionnel ses aventures au pays du libre accès[15]. Un *Datathon de la parole,* organisé pour encourager le dépôt, l'archivage et la diffusion de corpus oraux (linguistique, socio-linguistique, anthropologie, histoire orale)[16], confirmait la justesse de l'observation de Monica Heintz au sujet des effets induits par la réflexion au sujet des données de la recherche : « Certains chercheurs se sont aperçus qu'ils avaient encore des chantiers en cours, des données structurées mais non déposées (…), qu'il y avait des formes de soutien insoupçonnées de la part des ingénieurs de recherche qu'on n'avait pas toujours su exploiter, etc. » (Heintz 2023, 244).

Entretenir un niveau minimum d'information juridique : une composante importante du soin apporté aux données de la recherche

Une leçon de l'étape de l'inscription de la collection Pangloss au registre du CNRS, c'est qu'il importait, pour prendre soin[17] efficacement des données de la recherche, d'acquérir un niveau minimum d'information juridique. Ce point de vue est solidaire de l'idée (suggérée par Lionel Maurel) selon laquelle, la loi européenne ayant fait le choix d'un niveau élevé de protection des personnes (données personnelles, vie privée…), satisfaire aux obligations légales (dans le respect de l'esprit des lois) est en soi un objectif exigeant, au point qu'il n'est pas évident qu'il soit nécessaire de lui ajouter une « surcouche » d'éthique formelle. Cette perspective paraît tout à fait

12 « Je tâcherai d'allier toujours, dans cette recherche, ce que le droit permet avec ce que l'intérêt prescrit, afin que la justice et l'utilité ne se trouvent point divisées. » Jean-Jacques Rousseau, *Du contrat social, ou Principes du droit politique*, livre premier.

13 Voir notamment : https://userblogs.fu-berlin.de/langsci-press/2018/07/11/what-it-means-to-be-open-and-community-based-the-unicode-cookbook-as-a-showcase/

14 Cocoon, pour *COllections de COrpus Oraux Numériques*, est une plate-forme hébergeant plus de trente collections d'enregistrements de parole (y compris les langues des signes), principalement dédiées à la recherche et la médiation scientifique, plutôt qu'à des usages commerciaux. Voir : https://cocoon.huma-num.fr/

15 « Au pays des merveilles du libre accès : la préparation d'un premier ouvrage chez Language Science Press », Carreau de la BULAC (carnet de recherche de la Bibliothèque universitaire des langues et civilisations), 2017. https://doi.org/10.58079/m5ki

16 « Datathon de la parole, 8-10 novembre 2021 : dépôt, archivage et diffusion de documentation linguistique sur langues rares ». Les Carnets du LACITO. https://doi.org/10.58079/qpc8 Deuxième édition en 2023 au DataLab de la Bibliothèque nationale de France (au sujet duquel on consultera la présentation de Carlin et Laborderie 2021).

17 L'expression « prendre soin » est un clin d'œil à l'ouvrage *Prendre soin : de l'informatique et des générations* (Alombert et al. 2021), en hommage à Bernard Stiegler.

centrale pour le déploiement d'un modèle alternatif des comités d'éthique de la recherche, dans lesquels une place centrale soit accordée à la réflexivité (Gagnon 2010). Dans la recherche d'« un juste équilibre entre cette réflexivité et les nécessaires procédures à mettre en place » (Bazin et Goiseau 2023, 73), les exigences juridiques sont incontournables et doivent être intégrées (Bazin et Goiseau 2023, 93). En répondant à ces exigences de façon claire, nos disciplines se trouvent bien placées pour défendre leur point de vue, contre une bureaucratisation de l'éthique de la recherche.

Le recueil d'un consentement éclairé, dont la section « Autorisation de diffusion et consentement oral : un essai peu concluant » soulignait la difficulté, apparaît également sous un jour nettement moins problématique lorsqu'il est abordé avec l'appui de spécialistes des questions juridiques. Céline Aires, Déléguée à la protection des données de l'Université Sorbonne Nouvelle, propose, en concertation avec le réseau des Délégués à la protection des données, le principe d'un consentement éclairé oral devant témoin lettré. Ce témoin est une personne de confiance (autre que le chercheur), qui atteste par écrit que le consentement éclairé a bien été donné après une information adaptée au contexte culturel : que l'explication fournie a été comprise et acceptée. De la sorte, le chercheur n'est pas seul à décider que le participant est éclairé.

Grâce aux conseils avisés des collègues du CNRS et de ses institutions partenaires, la collection Pangloss semble parée au plan juridique (aussi bien qu'au plan technique et organisationnel) pour faire face à une montée en charge, et destinée à connaître un bel avenir. Dans ce contexte, les avancées rapides du Traitement Automatique des Langues suscitent de grands espoirs mais aussi des interrogations et des inquiétudes.

Traitement automatique des langues rares : un potentiel considérable pour la documentation linguistique… et des craintes de vol à l'étalage

Le caractère stratégique des outils informatiques pour les sciences humaines et sociales (Delmas-Rigoutsos 2023) est particulièrement manifeste dans le cas des sciences du langage. La perspective d'un soutien du Traitement Automatique des Langues (TAL) à la documentation linguistique constitue une lueur d'espérance, dans un contexte de déclin rapide de la diversité linguistique mondiale, parallèle au déclin de la biodiversité. Le fort potentiel applicatif du TAL pour la documentation des langues est bien identifié (Anastasopoulos *et al.* 2020), et des réalisations viennent peu à peu le concrétiser (Harrigan *et al.* 2023). L'alignement texte-parole permet un accès facilité aux ressources (Littell *et al.* 2022), et les outils de Reconnaissance Automatique de la Parole facilitent la transcription de documents audio et vidéo dans des langues rares, en mode *transcription exhaustive* (Partanen, Hämäläinen et Klooster 2020 ; Liu, Spence et Prud'hommeaux 2022 ; Guillaume *et al.* 2022) ou en mode *fouille de documents* : recherche de mots-clefs (Hjortnæs, Partanen et Tyers 2021). La synthèse de la parole trouve sa place dans des projets de revitalisation (Pine *et al.* 2022), de même que toutes sortes d'autres outils, par exemple pour le traitement des paradigmes verbaux de langues polysynthétiques (Kuhn *et al.* 2020). Les collaborations entre TAListes et spécialistes de langues rares ouvrent d'importantes perspectives en recherche, du fait des défis spécifiques posés par le scénario « à faibles ressources » que constitue le travail sur des langues peu documentées (Jimerson, Liu et Prud'Hommeaux 2023 ; Lonergan *et al.* 2023 ; Fily *et al.* 2024 ; San *et al.* 2024).

L'historique des collaborations *linguistique de terrain + TAL* nouées depuis 2014 autour de la transcription automatique de la langue na de Yongning a fait l'objet d'une relation circonstanciée à destination d'un public de linguistes de terrain (Michaud *et al.* 2018), complémentaire des articles de TAL au sujet du volet informatique (Do, Michaud et Castelli 2014 ; Adams *et al.* 2018 ; Wisniewski, Guillaume et Michaud 2020). Les présentes réflexions, qui s'organisent autour du thème de l'éthique de la recherche, fournissent l'occasion de relever la place que tenaient les questions d'éthique dans ces collaborations.

Éthique de la recherche et collaborations pluridisciplinaires

Un point central qui ressort des réflexions des équipes pluridisciplinaires « TAListes + linguistes » qui travaillent sur des langues rares est qu'une convergence au sujet d'une échelle de valeurs, dont découlent des objectifs partagés, est indispensable à des collaborations véritablement fructueuses. Si l'éthique est une réflexion sur les valeurs qui orientent et motivent nos actions, la valeur accordée à la documentation linguistique constitue un *ethos* – un lieu où s'établir et vivre une collaboration. Dès lors qu'on considère la constitution d'une documentation fiable et abondante sur une langue en danger comme une fin en soi, on pourra travailler à cette fin en bonne intelligence avec des collègues d'autres disciplines et d'autres métiers : informaticien·nes, enseignant·es, linguistes, archivistes, anthropologues, ethno-botanistes et d'autres encore.

> In the same way as some linguists feel more strongly than others about the value of language diversity, and come to identify language documentation and language description as priorities, some computer scientists consider language processing for under-resourced languages as their field of specialization. (…) The sense of a common goal fosters mutual interest between linguists and computer scientists, itself conducive to mutual understanding. (Michaud *et al.* 2018, 403)

Le caractère positif des expériences menées, ainsi que des utilisations non anticipées des données de la collection Pangloss par des chercheurs d'autres institutions (Guzmán *et al.* 2017 ; Flamein et Eshkol-Taravella 2021), nous amenait à encourager leur multiplication, notamment en facilitant l'utilisation de corpus de langues rares par des TAListes (Galliot *et al.* 2021). Le risque d'abus était connu ; mais la mise en libre accès de l'ensemble de la collection Pangloss avait également constitué une prise de risque, et vingt ans plus tard, on constatait que les retours d'utilisatrices et utilisateurs (certes peu nombreux) étaient uniformément positifs, ce qui allait dans le sens d'encourager le choix d'aller de l'avant. Pourtant, il est peu à peu apparu que les progrès du TAL n'amenaient pas seulement des collaborations passionnantes.

Incertitudes liées à l'évolution technologique rapide en Traitement Automatique des Langues

Le rythme rapide d'évolution du domaine du TAL fait qu'il n'est pas possible de prévoir toutes les réutilisations qui pourront être faites des données dans un avenir proche (Badin *et al.* 2022, §52). Cette observation est aussi déconcertante qu'essentielle. Il est certes possible de limiter les usages autorisés, par le choix judicieux de la licence sous laquelle sont placées les ressources orales. Mais l'expérience récente (dans les années 2020) de l'ingestion massive de données par les entreprises du numérique (pour l'entraînement de modèles statistiques) montre que les grandes entreprises ne sont pas toujours des acteurs respectueux des licences. L'Union Européenne affiche de saines ambitions (« Parliament's priority is to make sure that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly »[18]), mais leur opérationnalisation pourrait se faire attendre, d'autant que la rapidité de l'évolution technologique complique son encadrement juridique. Les dérives d'un modèle socio-économique prédateur de données (vol à l'étalage généralisé de données, et déploiement d'outils de TAL hors du contrôle des sociétés concernées) viennent aujourd'hui remettre en question le choix d'un accès ouvert aux données de langues rares. En 2018, un *keynote speaker* en sécurité informatique mettait en garde contre le déploiement précipité d'outils statistiques dits d'intelligence artificielle, et opposait, aux sirènes du solutionnisme technologique, la recommandation d'écouter la voix des poètes (Mickens 2018). En 2021, année où l'*Association for Computational Linguistics* mettait en place son Comité d'éthique, une étude relevait les dangers de la course au gigantisme dans les modèles de langues. Outre les dégâts environnementaux et la reproduction de biais sociaux, le

18 « EU AI Act : first regulation on artificial intelligence », https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

risque de graves malentendus est inhérent au fait d'exposer des humains à un message *qui n'a été produit par personne*. « [T]he human tendency to attribute meaning to text, in combination with large language models' ability to learn patterns of forms that humans associate with various biases and other harmful attitudes, leads to risks of real-world harm, should language-model-generated text be disseminated » (Bender *et al.* 2021, 618). Ces observations ne donnent pas tort aux formulations (certes plus vagues) d'un de nos poètes, qui écrivait, cinq ans avant sa disparition :

> Nous sommes des milliards à tourbillonner, derviches, de plus en plus fortement, étourdis, vertigineux, solipsistes, reclus en notre *self*. (…) le tournis général … donné en spectacles innombrables de télé-vérité à tous les écrans de notre « vivre en direct » : équivalemment milliards d'opinions horaires à somme nulle, établissant le régime *« post-truth/non-truth »*, disons « trumpiste », de l'oralité humaine : ce sont nos *big data*, le mauvais infini, que les autres, les Big Data scientifiques, engloutiront. (Michel Deguy, *L'envergure des comparses* (2017), Paris : Hermann, pp. 39-40.)

Pour revenir aux responsabilités des linguistes qui collectent des données sur le terrain, l'époque de l'archivisme insouciant est bien révolue. On hésite à laisser en accès public des corpus qui pourraient accélérer la généralisation aux langues les plus rares des outils de surveillance de masse désormais bien implantés dans nos sociétés. La voix de la prudence peut conseiller d'éviter l'ouverture de corpus qui pourraient faire l'objet d'utilisations non souhaitables.

Conclusion : Pangloss et la Reine Rouge, ou le jardinage et son éthique

Au fil des réflexions, il ressort que l'adoption conséquente et résolue de principes de Science ouverte ne constitue pas par elle-même une réponse complète (et encore moins une réponse pérenne, *future-proof*) aux questions juridiques et éthiques soulevées par la constitution, la publication électronique et l'exploitation de ressources linguistiques en langues rares. Les enjeux se déplacent d'une recherche de solutions simples et claires qui permettent d'éviter les soucis (on fait ce qu'il faut pour satisfaire aux critères en vigueur en matière d'éthique, d'ouverture des données, de partage des outils, de libre accès aux publications, et on a l'esprit libre pour se concentrer sur la recherche) vers un processus réflexif continu. Une réflexion exigeante et vigilante, naturellement soucieuse de respect des personnes et des groupes sociaux, informe l'ensemble du processus de documentation et de recherche, du stade de la définition des protocoles d'enquête jusqu'aux choix concernant la gestion des données sur le moyen terme et le long terme.

La référence au maître Pangloss de Candide (non dénuée d'autodérision, quand on connaît la nouvelle de Voltaire) oriente la collection Pangloss vers une certaine conception de l'éthique : celle d'une petite équipe qui cultive son jardin, *les pieds sur terre*, se défiant des approches et préconisations « hors sol ». Le jardin de données qu'est la collection Pangloss, conçu comme un Commun de la connaissance en accès ouvert, doit-il devenir secret, pour se protéger contre des usages contraires à toute éthique ? Rien ne dit que la tentation du repli soit bonne conseillère. La métaphore du jardinage, qui souligne le caractère continu de l'activité de maintenance et d'amélioration de la collection, suggère aussi que l'archive est, comme le vivant, prise dans une évolution permanente qui impose d'être constamment mobile pour parvenir à rester en place (courir tant qu'on peut pour rester au même endroit, dit la Reine Rouge dans *De l'autre côté du miroir*). L'adaptation est toujours à recommencer, et l'extinction toujours possible.

Remerciements

Mes remerciements aux professionnels des archives orales qui m'ont fait découvrir leur métier au fil des échanges : en particulier Michel Jacobson, Joséphine Simonnot, Florence Gétreau, Pascal

Cordereix. Merci à l'équipe de la collection Pangloss (Séverine Guillaume, Léa Mouton, Balthazar Do Nascimento) pour son appui constant, et à Flora Badin, Natalia Cáceres, Raphaëlle Chossenot et Julie Giovacchini pour leur vision stratégique et leurs conseils judicieux.

Merci aux organisatrices de la Journée *Éthique et TAL 2024* de l'Association pour le Traitement Automatique des Langues (ATALA), Karën Fort et Aurélie Névéol, dont la sollicitation a fourni l'occasion du présent travail.

Merci à Aliyah Morgenstern, Julie Marsault, Séverine Guillaume, Léa Mouton et Marc-Antoine Mahieu pour leur relecture, et à Evangelia Adamou et Jacqueline Vaissière pour leurs réflexions au fil de nos échanges.

Il va de soi que les points de vue exposés ici n'engagent que moi.

Références citées

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird et Alexis Michaud. 2018. « Evaluating phonemic transcription of low-resource tonal languages for language documentation ». Dans *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3356-3365. Miyazaki. https://halshs.archives-ouvertes.fr/halshs-01709648

Alombert, Anne, Victor Chaix, Maël Montévil et Vincent Puig, éd. 2021. *Prendre soin: de l'informatique et des générations*. Limoges : FYP.

Anastasopoulos, Antonios, Christopher Cox, Graham Neubig et Hilaria Cruz. 2020. « Endangered languages meet Modern NLP ». Dans *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, 39-45.

Archambault, Éric et Vincent Larivière. 2009. « History of the journal impact factor: Contingencies and consequences ». *Scientometrics* 79 (3) : 635-649.

Arènes, Cécile, Lionel Maurel et Stéphanie Rennes. 2022. « Guide d'application de la Loi pour une République numérique pour les données de la recherche ». Ministère de l'enseignement supérieur et de la recherche. https://doi.org/10.52949/31.

Badin, Flora, Caroline Cance, Céline Dugua, Layal Kanaan-Caillol, Anne-Lyse Minard et Katja Ploog. 2022. « Les données orales en linguistique: Questions éthiques et cadre juridique ». *Bulletin de l'Association Française d'Archives Sonores (AFAS)*, nº 48 (décembre) : 158-181. https://doi.org/10.4000/afas.7496

Bazin, Yoann et Élise Goiseau. 2023. « Vers un modèle alternatif des comités d'éthique de la recherche : Quel équilibre entre procédures et réflexivité ? » *Revue française de gestion* 49 (1). Cairn/Isako : 73-100.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major et Shmargaret Shmitchell. 2021. « On the dangers of stochastic parrots: can Language Models be too big? ». Dans *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.

Bendjaballah, Selma et Guillaume Garcia. 2023. « Les sciences sociales à l'épreuve de l'ouverture des données de la recherche ». *Terrains & travaux* 43 (2). Cachan : ENS Paris-Saclay : 211-216. https://doi.org/10.3917/tt.043.0211

Bichurina, Natalia. 2017. « Baptêmes d'une langue ou un peu de magie sociale dans le passé et dans le présent (francoprovençal–arpitan–savoyard) ». *Cahiers du Centre de Linguistique et des Sciences du Langage* 52 : 119-138.

Bonnemason, Bénédicte, Véronique Ginouvès et Véronique Pérennou. 2001. *Guide d'analyse documentaire du son inédit, pour la mise en place de banques de données*. Parthenay, France : Modal - AFAS.

Bordignon, Frédérique. 2023. « Maintenir l'intégrité de la littérature scientifique à l'heure de l'ouverture de la science : étude de cas, enjeux techniques et rôle des acteurs | 3e session de l'Assemblée des partenaires de HAL ». Video/mp4. Centre pour la Communication Scientifique Directe. https://doi.org/10.60527/7DMS-H251

Bouquiaux, Luc et Jacqueline Thomas. 1971. *Enquête et description des langues à tradition orale. Volume I: l'enquête de terrain et l'analyse grammaticale*. 2nd edition 1976. Paris : Société d'études linguistiques et anthropologiques de France.

Bowern, Claire. 2010. « Fieldwork and the IRB: A snapshot ». *Language* 86 (4). Linguistic Society of America : 897-905.

Caplan, Patricia, éd. 2003. *The ethics of anthropology: debates and dilemmas*. London; New York : Routledge.

Carlin, Marie et Arnaud Laborderie. 2021. « Le BnF DataLab, un service aux chercheurs en humanités numériques ». *Humanités numériques*, n° 4. https://doi.org/10.4000/revuehn.2684

Cox, Christopher. 2022. « Managing Data in a Language Documentation Corpus ». Dans *The Open Handbook of Linguistic Data Management*, édité par Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller et Lauren B. Collister, 277-286. The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0027

Delmas-Rigoutsos, Yannis. 2023. « Numérique et humanités : de l'ancillarité à la fécondité grâce à la modélisation computationnelle des connaissances ». *Humanités numériques* 7. https://doi.org/10.4000/revuehn.3359

Do, Thi Ngoc Diep, Alexis Michaud et Eric Castelli. 2014. « Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages ». Dans *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, 153-160. St Petersburg. http://halshs.archives-ouvertes.fr/halshs-00980431

Driem, George van. 2016. « Endangered language research and the moral depravity of ethics protocols ». *Language Documentation and Conservation* 10 : 243-252.

Eslen-Ziya, Hande, Alberta Giorgi et Ceren J. Ahi. 2023. « Digital Vulnerabilities and Online Harassment of Academics, Consequences, and Coping Strategies. An Exploratory Analysis ». *Feminist Media Studies*, novembre, 1-6. https://doi.org/10.1080/14680777.2023.2281268

Fassoulaki, A., K. Papilas, A. Paraskeva et K. Patris. 2002. « Impact factor bias and proposed adjustments for its determination ». *Acta Anaesthesiologica Scandinavica* 46 (7). Wiley Online Library : 902-905.

Ferlus, Michel. 2017. « Des enregistrements sans expérience à la numérisation : les insouciances du terrain ». *La lettre de l'AFRASE* 93-94 : 12-13.

Fily, Maxime, Guillaume Wisniewski, Severine Guillaume, Gilles Adda et Alexis Michaud. 2024. « Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models ». Dans *Findings of the Association for Computational Linguistics (EACL 2024)*. Malta. http://arxiv.org/abs/2402.05581

Flamein, Hélène et Iris Eshkol-Taravella. 2021. « Exploitation du corpus Enquêtes sociolinguistiques à Orléans (ESLO) par les outils du traitement automatique des langues et de la géomatique ». *Humanités numériques*, nº 3. https://doi.org/10.4000/revuehn.1911

Fluehr-Lobban, Carolyn. 2013. *Ethics and Anthropology: Ideas and Practice*. Lanham, MD : AltaMira press.

Gagnon, Éric. 2010. « Le comité d'éthique de la recherche, et au-delà ». *Éthique publique* 12 (1) : 299-308. https://doi.org/10.4000/ethiquepublique.284

Galliot, Benjamin, Guillaume Wisniewski, Séverine Guillaume, Alexis Michaud, Solange Rossato, Minh-Châu Nguyên et Maxime Fily. 2021. « Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal ». Dans *Journées scientifiques du Groupement de recherche « Linguistique informatique, formelle et de terrain » (GDR LIFT)*. Grenoble. https://halshs.archives-ouvertes.fr/halshs-03475436

Galonnier, Juliette, Stefan Le Courant, Anthony Pecqueux et Camille Noûs. 2019. « Ouvrir les données de la recherche ? » *Tracés*, nº 19 : 17-33. https://doi.org/10.4000/traces.10588

Garellek, Marc, Matthew Gordon, James Kirby, Wai-Sum Lee, Alexis Michaud, Christine Mooshammer, Oliver Niebuhr, *et al.* 2020. « Toward Open Data Policies in Phonetics: What We Can Gain and How We Can Avoid Pitfalls ». *Journal of Speech Science* 9 (1). https://halshs.archives-ouvertes.fr/halshs-02894375

Gros, Stéphane. 2014. « The bittersweet taste of rice. Sloping land conversion and the shifting livelihoods of the Drung in Northwest Yunnan (China) ». *Himalaya* 34 (2) : 81-96.

Guillaume, Séverine, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyễn, Maxime Fily, Guillaume Jacques et Alexis Michaud. 2022. « Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings ». Dans *Proceedings of Interspeech 2022*. Incheon, Korea. https://halshs.archives-ouvertes.fr/halshs-03625581

Gunthert, André. 2004. « La punition des revues ». *Études photographiques*, nº 15. Société française de photographie : 2-3.

Guzmán, Gualberto A., Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock et Almeida Jacqueline Toribio. 2017. « Metrics for modeling code-switching across corpora ». Dans *Proceedings of Interspeech 2017*, 67-71.

Harrigan, Atticus, Aditi Chaudhary, Shruti Rijhwani, Sarah Moeller, Antti Arppe, Alexis Palmer, Ryan Henke et Daisy Rosenblum. 2023. *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-6)*. Association for Computational Linguistics (ACL) Anthology.

Harris, Amanda, Nick Thieberger et Linda Barwick. 2015. *Research, records and responsibility: Ten years of PARADISEC*. Sydney University Press.

Heintz, Monica. 2023. « Protéger ou invisibiliser ses interlocuteurs : peut-on ouvrir les données ethnographiques ? » Édité par Selma Bendjaballah et Guillaume Garcia. *Terrains & travaux* 43 (2). Cachan : ENS Paris-Saclay : 233-255. https://doi.org/10.3917/tt.043.0211

Henitiuk, Valerie et Marc-Antoine Mahieu. 2024. « Tangled lines: what might it mean to take Indigenous languages seriously? » *Translation Studies* 17 (1). Taylor & Francis : 169-180.

Hjortnæs, Nils, Niko Partanen et Francis Tyers. 2021. « Keyword spotting for audiovisual archival search in Uralic languages ». Dans *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, 1-7. Association for Computational Linguistics.

Jacobson, Michel, Boyd Michailovsky et John B. Lowe. 2001. « Linguistic documents synchronizing sound and text ». *Speech Communication* 33 [special issue: "Speech Annotation and Corpus Tools"] : 79-96.

Jimerson, Robert, Zoey Liu et Emily Prud'Hommeaux. 2023. « An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language ». Dans *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1008-1016.

Kaufmann, Sina Kamala, Michael Timmermann, Annemarie Botzki et Steffen Greiner, éd. 2019. *Wann wenn nicht wir*: ein extinction rebellion Handbuch*. Traduit par Ulrike Bischoff. Erweiterte deutsche Erstausgabe. Frankfurt am Main : S. Fischer.

Kuhn, Roland, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine et Caroline Running Wolf. 2020. « The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software ». Dans *Proceedings of the 28th international conference on computational linguistics*, 5866-5878.

Launey, Michel. 2023. *La République et les langues*. Cours et travaux. Paris : Raisons d'agir.

Littell, Patrick, Eric Joanis, Aidan Pine, Marc Tessier, David Huggins-Daines et Delasie Torkornoo. 2022. « Readalong studio: Practical zero-shot text-speech alignment for indigenous language audiobooks ». Dans *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, 23-32.

Liu, Zoey, Justin Spence et Emily Prud'hommeaux. 2022. « Enhancing documentation of Hupa with automatic speech recognition ». Dans *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Lonergan, Liam, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl et Ailbhe Ní Chasaide. 2023. « Towards Spoken Dialect Identification of Irish ». Dans *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-Resourced Languages (SIGUL 2023)*, 63-67. ISCA. https://doi.org/10.21437/SIGUL.2023-14

Marsault, Julie. 2021. « Valency-Changing Operations in Umó^{n}ho^{n} : Affixation, Incorporation, and Syntactic Constructions, Les Modifications de La Valence Verbale En Umó^{n}ho^{n} : Affixation, Incorporation et Constructions Syntaxiques ». Thèse de doctorat, Université de la Sorbonne nouvelle - Paris III. https://theses.hal.science/tel-03573762

Michaud, Alexis. 2002. « "Tu pourrais enregistrer un corpus pour moi?" Pour une charte de qualité des corpus ». Dans *XXIVe Journées d'Etude de la Parole*, 153-156. Nancy, France. https://shs.hal.science/halshs-01647020

Michaud, Alexis. 2017a. « In memoriam He Qin (1973-2016) — 纪念和沁教授 ». Text/html. *Indo-Sinica · 震南古聲隨記*. https://doi.org/10.58079/Q69H

Michaud, Alexis. 2017b. « Le projet de numérisation DO-RE-MI-FA : données des recherches de Michel Ferlus en Asie du Sud-Est ». *Lettre de l'AFRASE (Association Française pour la Recherche en Asie du Sud-Est)*, 2017.

Michaud, Alexis. 2017c. *Tone in Yongning Na: lexical tones and morphotonology*. Studies in Diversity Linguistics 13. Berlin : Language Science Press. http://langsci-press.org/catalog/book/109

Michaud, Alexis, Oliver Adams, Trevor Cohn, Graham Neubig et Séverine Guillaume. 2018. « Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit ». *Language Documentation & Conservation* 12 : 393-429.

Michaud, Alexis, Minh-Châu Nguyễn et He Likun 和丽昆. 2020. « Voix de « ceux qui ne sont rien » en Asie du Sud-Est ». *Cahiers de littérature orale*, nº Hors-Série : 147-154. https://doi.org/10.4000/clo.7019

Mickens, James. 2018. « Q: Why do keynote speakers keep suggesting that improving security is possible? A: Because keynote speakers make bad life decisions and are poor role models ». Dans *27th USENIX Security Symposium (USENIX Security 18)*.

Middeldorp, Nick et Philippe Le Billon. 2020. « Deadly environmental governance: authoritarianism, eco-populism, and the repression of environmental and land defenders ». Dans *Environmental governance in a populist/authoritarian era*, 24-37. Routledge.

Partanen, Niko, Mika Hämäläinen et Tiina Klooster. 2020. « Speech recognition for endangered and extinct Samoyedic languages ». Dans *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. https://arxiv.org/abs/2012.05331

Pine, Aidan, Dan Wells, Nathan Brinklow, Patrick Littell et Korin Richmond. 2022. « Requirements and motivations of low-resource speech synthesis for language revitalization ». Dans *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7346-7359.

Ripple, William J., Christopher Wolf, Thomas M. Newsome, Mauro Galetti, Mohammed Alamgir, Eileen Crist, Mahmoud I. Mahmoud, William F. Laurance et 364 scientist signatories from 184 countries 15. 2017. « World scientists' warning to humanity: a second notice ». *BioScience* 67 (12) : 1026-1028.

Rooryck, Johan. 2016. « Introducing Glossa ». *Glossa* 1 (1). Ubiquity Press : 1-3. https://doi.org/10.5334/gjgl.91

San, Nay, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams et Dan Jurafsky. 2024. « Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens ». *arXiv preprint arXiv:2402.02302*.

Schrag, Zachary M. 2010. *Ethical Imperialism: Institutional Review Boards and the Social Sciences 1965-2009*. Baltimore, MD : John Hopkins University Press.

Simon, Camille et Camille Noûs. 2021. « The grammaticalization of plurality in the languages of Amdo ». *Himalayan Linguistics* 20 (3) : 49-81.

Simonnot, Joséphine. 2020. « Partager les archives sonores du musée de l'Homme sur le web avec la plateforme Telemeta ». *Bulletin de l'Association Française d'Archives Sonores (AFAS)* 46 : 88-101. https://doi.org/10.4000/afas.4056

Simons, Gary et Steven Bird. 2000. « The seven pillars of open language archiving: A vision statement ». *Open Language Archives Community. Available at http://www.language-archives.org/docs/vision.html*. 2000.

Snead, Taylor et Ellen Cushman. 2023. « Building a community-centered archive for Cherokee language description, documentation, and reclamation ». *The Modern Language Journal* 107 (1). Wiley Online Library : 242-267.

Vapnarsky, Valentina. 2020. « Retour aux sources ? Circulation et virtualités des savoirs amérindiens à l'ère du numérique ». *Journal de la société des américanistes* 106 (2) : 79-103. https://doi.org/10.4000/jsa.19003

Vasile, Aurelia, Séverine Guillaume, Mourad Aouini et Alexis Michaud. 2020. « Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger ». *I2D - Information, données & documents* 2 : 156-175.

Vernaudon, Jacques, Nick Thieberger, Tamatoa Bambridge et Takurua Parent. 2021. « Un nouveau souffle numérique pour les corpus en langues océaniennes ». *Journal de la société des océanistes*, nº 153 (décembre) : 323-336. https://doi.org/10.4000/jso.13165

Wisniewski, Guillaume, Séverine Guillaume et Alexis Michaud. 2020. « Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? » Dans *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, édité par Dorothee Beermann, Laurent Besacier, Sakriani Sakti et Claudia Soria, 306-315. Marseille, France : European Language Resources Association (ELRA). https://halshs.archives-ouvertes.fr/hal-02513914/

Woodbury, Tony. 2011. « Language documentation ». Dans *The handbook of endangered languages*, édité par Peter Austin et Julia Sallabank, 1 : 35-51. Cambridge : Cambridge University Press.