HAL open science

Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo

Alexis Michaud, Séverine Guillaume, Guillaume Jacques, Đăng-Khoa Mạc, Michel Jacobson, Thu-Hà Phạm, Matthew Deo

▶ **To cite this version:**

Alexis Michaud, Séverine Guillaume, Guillaume Jacques, Đăng-Khoa Mạc, Michel Jacobson, et al.. Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. Journées d'Etude de la Parole 2016, Association Francophone de la Communication Parlée, Jul 2016, Paris, France. pp.155-163. halshs-01341631

HAL Id: halshs-01341631

https://shs.hal.science/halshs-01341631v1

Submitted on 4 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo

Alexis Michaud[1, 2] Séverine Guillaume[1] Guillaume Jacques[3] Đăng-Khoa Mạc[2]
Michel Jacobson[4] Thu-Hà Phạm[5] Matthew Deo
(1) Langues et Civilisations à Tradition Orale, CNRS - Sorbonne Nouvelle - Institut national des langues et civilisations orientales, Paris, France
(2) Institut international de recherche MICA, Hanoi University of Science and Technology - CNRS - Grenoble INP, Hanoi, Vietnam
(3) Centre de Recherches Linguistiques sur l'Asie Orientale, CNRS - Ecole des Hautes Etudes en Sciences Sociales, Paris, France
(4) Laboratoire Ligérien de Linguistique, Universités d'Orléans et de Tours - BnF - CNRS
(5) Department of Linguistics, VNU University of Social Sciences and Humanities, Hanoi

alexis.michaud@vjf.cnrs.fr, severine.guillaume@vjf.cnrs.fr, rgyalrongskad@gmail.com, dang-khoa.mac@mica.edu.vn, michel.jacobson@gmail.com, phamha.ling@gmail.com, matthewdeo@gmail.com

RESUME

La présente communication présente les projets scientifiques et les réalisations de deux collections hébergées par la plateforme de ressources orales Cocoon : la Collection Pangloss, qui concerne principalement des langues de tradition orale (sans écriture), du monde entier ; et la Collection AuCo, dédiée aux langues du Vietnam et de pays voisins. L'objectif est un progrès solidaire des recherches et de la documentation linguistique. L'accent est mis sur les perspectives ouvertes pour la recherche en phonétique/phonologie par certaines réalisations récentes dans le cadre de ces deux Collections.

ABSTRACT

Contributing to joint progress in documentation and research: some achievements and future perspectives of the Pangloss Collection and the AuCo Collection

This talk sets out the scientific goals and achievements of two collections hosted by the Cocoon Open Archive of oral resources: the Pangloss Collection, which mainly focuses on unwritten languages from all areas in the world ; and the AuCo Collection, which is dedicated to languages of Vietnam and neighbouring countries. The aim is to contribute to joint progress in language documentation and in research. Emphasis is placed on the perspectives for phonetic/phonological research that are opened by some recent achievements in the framework of these two Collections.

MOTS-CLES : recherches phonétiques ; documentation linguistique ; archives orales ; archives ouvertes ; langues peu dotées ; diversité linguistique ; documentation en danger.

KEYWORDS: phonetic research; language documentation; language archives; open archives; under-resourced languages; linguistic diversity; endangered documentation.

1 Introduction

La présente communication présente les projets scientifiques et les réalisations de deux collections hébergées par la plateforme de ressources orales Cocoon : la Collection Pangloss, qui concerne principalement des langues de tradition orale (sans écriture), du monde entier ; et la Collection AuCo, dédiée aux langues du Vietnam et de pays voisins. L'objectif est un progrès solidaire des recherches et de la documentation linguistique. L'accent est mis sur les perspectives ouvertes pour la recherche en phonétique/phonologie par certaines réalisations récentes dans le cadre de ces deux Collections.

1.1 Etat des lieux : la prise de données à usage unique demeure pratique courante

Les bases empiriques des recherches phonétiques demeurent à l'heure actuelle un point de fragilité. La prise de données constitue un défi souvent sous-estimé (Niebuhr & Michaud 2015). Les bases de données sonores des centres de recherches en phonétique restent paradoxalement assez peu structurées, et relativement peu employées. Chercheurs et étudiants ont souvent tendance à constituer leur propre corpus en fonction des besoins de leur recherche, considérant qu'il est plus commode de recueillir de nouvelles données que de réemployer des ensembles documentaires existants. De fait, les recherches en phonétique nécessitent des données qui répondent à des critères précis concernant notamment les locuteurs et le type de méthodes expérimentales, critères que les données recueillies pour le propos d'expériences antérieures peuvent rarement satisfaire en intégralité. Les fonds d'archives restent relativement peu connus dans les laboratoires de phonétique. Les grands corpus distribués sur Internet peuvent être trop coûteux pour des recherches fondamentales (sans application commerciale directe) ; or les technologies numériques permettent d'enregistrer soi-même des données facilement et à faible coût. Des outils logiciels comme SpeechRecorder[1] permettent en outre de simplifier considérablement le travail d'édition et d'annotation des enregistrements.

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, et des activistes de la documentation soulignent depuis des années l'importance de la conservation et la diffusion des données (voir notamment Thieberger et al. 2016). Mais dans l'étude des grandes langues, la prise de données gagnerait également à être conçue dans une logique de progrès cumulatif de la documentation, au lieu de collecter des données à usage unique (dont le réemploi n'est pas prévu d'emblée). Quiconque a l'expérience de la collecte de données confirmera qu'il s'agit d'une activité qui s'avère chronophage au final. Les étapes sont nombreuses : mise au point du protocole expérimental, tests, rendez-vous avec les participants, enregistrement, mise en forme... L'absence de réutilisation des données constitue une déperdition pour la communauté des recherches en parole.

Si l'on enregistre des données en ayant en tête la perspective d'un archivage pérenne et d'un partage auprès de la communauté, cela encourage à accroître l'investissement initial de temps et de soin, ce qui a, selon notre expérience, des conséquences positives en termes de qualité des données, et partant, de fiabilité des recherches. Pour ne prendre qu'un exemple, celui du choix des locuteurs : les usagers des laboratoires de phonétique sont souvent sollicités comme sujets pour enregistrer des

[1] http://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

données. Ils ont l'expérience des tâches demandées, tandis que les non-initiés peuvent être intimidés ou perplexes ; par ailleurs, étudiants et collègues peuvent rendre service bénévolement. Mais le fait de recourir à un locuteur linguiste, souvent polyglotte, pose des problèmes épistémologiques évidents. Des mots français enregistrés par un locuteur natif pour un cours de lecture de spectrogrammes se sont avérés « non canoniques » au point d'induire en erreur des déchiffreurs chevronnés ; c'est vraisemblablement la conséquence d'une expérience linguistique diversifiée. Il n'y a rien de surprenant à ce que des locuteurs du japonais ou du vietnamien, après plusieurs mois en France, transposent dans leur langue maternelle les continuations intonatives qu'ils ont appris à employer en français (Dô Thê Dung, Trân Thien Huong & Boulakia 1998). Outre leur prononciation, ce séjour perturbe également leur façon de percevoir. Cela remet en cause certaines données publiées dans les revues internationales, fournies par des locuteurs natifs mais résidents depuis fort longtemps dans un pays étranger.

Ces fragilités restent actuellement masquées par le fait que les données ne sont généralement pas communiquées aux évaluateurs des travaux soumis aux revues scientifiques, ni aux lecteurs de ces travaux dans leur version publiée. En l'absence d'exigence de communication des données, que ce soit de la part des revues scientifiques ou des institutions qui financent le chercheur, pourquoi s'imposer des efforts supplémentaires, qui n'aideront pas à la publication des recherches, et ne compteront pas dans l'évaluation du chercheur ? Rendre ses données disponibles, c'est également prêter le flanc à la critique. Faudrait-il réécouter tous les enregistrements, pour vérifier qu'il n'y traîne pas un passage à écarter avant diffusion : fou-rires, raclements de gorge ou autres maladresses qui fourniraient matière à un montage audio tournant en ridicule les locuteurs et les auteurs ? La conclusion paraît claire : on a beaucoup à perdre (à commencer par un temps précieux) à partager ses données, et rien à y gagner.

Cette situation paradoxale n'a pas fondamentalement changé depuis un état des lieux présenté aux Journées d'Etude de la Parole il y a quatorze ans (Michaud 2002). On aimerait néanmoins essayer ici d'argumenter qu'on a beaucoup à gagner à introduire, dans sa pratique de recherche, un raisonnement en termes de progrès cumulatif de la documentation linguistique, solidaire de progrès de la recherche.

1.2 Problématique : associer documentation et recherche, pour leur bénéfice mutuel

Est-ce un hasard si les centres de recherche en phonétique qui mettent à disposition leurs collections sonores, sans être retenus par la crainte des critiques que pourraient attirer la qualité inégale des enregistrements, sont des pionniers mondiaux du domaine ? Le laboratoire de l'Université de Californie à Los Angeles a choisi de mettre en ligne des ressources abondantes, qui figurent en bonne place sur leur site internet (http://www.phonetics.ucla.edu/). Faut-il conclure que le temps consacré à l'archivage et la diffusion de données soit un luxe réservé aux chercheurs qui, forts d'un succès incontesté et d'une situation professionnelle assurée, peuvent se permettre des activités non rentables en termes de carrière ? Il nous paraît au contraire que le souci d'associer documentation et recherche est l'un des facteurs de la réussite du laboratoire de phonétique de UCLA. La valorisation du socle empirique de la recherche procède de la même logique qui a abouti au livre *The Sounds of the World's Languages* (Ladefoged & Maddieson 1996). L'une et l'autre de ces productions nous paraissent refléter la vision d'avenir de pionniers qui font le point des données et des analyses qu'ils ont dans leurs cartons, et qui publient cet inventaire de l'état de l'art, en le présentant pour ce qu'il est : une étape sur le chemin d'un progrès vers une compréhension plus complète de la face sonore des langues du monde.

2 La Collection Pangloss et ses usages pour la phonétique et le traitement automatique des langues

2.1 Présentation

La Collection Pangloss est une archive publique qui contient plus de 2.000 enregistrements (400 heures) en plus de 130 langues et dialectes, dont 990 documents annotés par une vingtaine de chercheurs. La Collection Pangloss réunit des documents linguistiques sonores, avec une spécialité de langues « rares » ou peu étudiées. Son but est de contribuer à la documentation et à l'étude du patrimoine humain que représentent les langues du monde.

La Collection Pangloss donne accès aux enregistrements sonores d'origine aussi bien qu'aux transcriptions et traductions ; c'est une garantie d'authenticité et une ressource pour la recherche. Les ressources associent donc son et texte. L'aspect texte comprend une transcription phonologique accompagnée, selon les cas, de représentations orthographiques (là ou celles-ci existent, y compris dans des écritures non latines), de traductions en diverses langues (généralement : anglais, français, et/ou langue nationale du pays d'enquête), de gloses morphologiques, de notes, etc. L'alignement de la transcription avec le son se fait généralement au niveau de la « phrase » ou du groupe intonatif, mais peut se faire également au niveau du mot ou du morphème.

La parole spontanée forme la plus grande partie du fonds : des documents enregistrés dans leur contexte social et transcrits en consultation avec les locuteurs. Mais la Collection contient également des séances d'enquête et des listes de mots, enregistrés et annotés par des chercheurs d'horizons très variés. La pérennité de ces ressources « rares » est assurée par Cocoon[2], archive structurée selon les normes actuelles (XML, OLAC, Dublin Core...), dans un format ouvert. Aussi bien ces données que les outils qui servent à leur préparation et leur diffusion sont librement disponibles sur le site de la Collection Pangloss[3]. Pour plus de détails concernant l'historique du projet, les choix technologiques et l'état actuel des collections, un article complet est disponible en ligne (Michailovsky et al. 2014) ; le présent exposé s'attache plus spécifiquement aux projets et collaborations possibles avec la communauté des phonéticiens/phonologues.

2.2 Usages pour l'enseignement et la recherche en phonétique

La Collection Pangloss se prête à divers usages pour l'enseignement et la recherche en phonétique ; usages qui, en retour, fournissent l'occasion d'un enrichissement.

Tout d'abord, on y trouve une illustration d'un grand nombre de sons des langues du monde. La langue oubykh, aujourd'hui éteinte, était l'une des deux langues les plus riches en consonnes jamais observées. Les enregistrements disponibles en ligne comportent diverses paires minimales enregistrées avec soin, qui ont servi à des tests de perception, et se prêtent aisément à une exploitation pour des enseignements en phonétique.

---

[2] *Collections de corpus oraux numériques* (Cocoon): http://cocoon.huma-num.fr Au sujet de cette plate-forme technique pour la gestion de collections de ressources orales, voir Jacobson et al. (2015), qui présente son fonctionnement ainsi que les innovations et expérimentations en cours, dont bénéficie l'ensemble des collections hébergées.

[3] http://lacito.vjf.cnrs.fr/pangloss/

Les données de la Collection Pangloss peuvent en outre fournir la matière à de nouvelles recherches au sujet de systèmes sonores. Pour reprendre l'exemple de l'oubykh, des enregistrements de cinéradiographie ont été réalisés, et ont donné lieu à une « Etude articulatoire de quelques sons de l'oubykh d'après film aux rayons X » (Leroy & Paris 1974). Les films aux rayons X ont été numérisés en 2001, mais n'ont pas encore été alignés avec les transcriptions, ni diffusés. Un travail de recherche au sujet de ces données, par exemple une étude comparée de certains types de consonnes dans diverses langues (pour le français: Bothorel et al. 1986), fournirait l'occasion de les préparer pour une mise en ligne, ce qui enrichirait l'ensemble documentaire.

Des applications en traitement automatique de la parole sont également envisageables. Les données de la Collection Pangloss comportent des annotations de grande qualité, fruit du travail de linguistes qui consacrent souvent la majeure partie de leur carrière à une langue ou un petit groupe de langues. Ces données, malgré leur faible volume, présentent par là un intérêt pour le traitement automatique : linguistique de corpus (concordances et étude des collocations), reconnaissance automatique de la parole, synthèse de la parole, ou traduction automatique.

A titre d'exemple : données de langue na de Yongning

Une communication soulignant la possibilité d'appliquer des traitements automatiques au corpus de langue na de Yongning (Michaud et al. 2012), assortie d'une étude-pilote en reconnaissance automatique (Do, Michaud & Castelli 2014), a attiré l'attention d'équipes travaillant en reconnaissance automatique, qui ont entrepris d'utiliser ces données de la Collection Pangloss pour des recherches d'avant-garde : tester les possibilités de parvenir à une transcription phonétique sans entraînement préalable. La transcription du linguiste sert de référence *(gold standard)* pour évaluer le degré de précision atteint par les algorithmes. On peut imaginer de nombreux autres scénarios, par exemple une collaboration entre chercheurs en informatique et en linguistique pour l'étude de la réalisation phonétique des séquences tonales. Un algorithme de reconnaissance automatique des tons serait entraîné au moyen du corpus déjà transcrit par le linguiste (environ cinq heures de parole). Un dialogue entre linguistes et informaticiens permettrait d'améliorer simultanément l'outil informatique et la modélisation linguistique, en testant l'emploi de paramètres proposés par le linguiste. Cela permettrait d'affiner les hypothèses au moyen d'une implémentation informatique. On pourrait par exemple imaginer de déterminer quelle proportion de l'information contenue dans la courbe de fréquence fondamentale peut s'expliquer par l'identité phonologique du ton (dans cette langue : Bas, Moyen, Haut, Bas-Haut, ou Moyen-Haut) ; par la déclinaison (schéma global d'évolution de la fréquence fondamentale au cours de l'énoncé) ; par le découpage en constituants ; et par la structure de l'information. Il pourrait être intéressant de montrer, par exemple, qu'en l'absence de prise en compte des paramètres contextuels (coarticulation tonale, déclinaison...), le pourcentage d'identification exacte n'est que de tant pour cent ; que la prise en compte de la coarticulation tonale améliore le système de tant pour cent ; que l'introduction d'une distinction entre mots pleins et mots-outils apporte une amélioration de tant, et ainsi de suite. A mesure des progrès de l'outil de reconnaissance développé pour la langue-cible, le résidu (tons non reconnus) serait de plus en plus limité, et de plus en plus intéressant pour une analyse qualitative : le linguiste pourrait exercer sa sagacité sur les passages qui induisent en erreur les algorithmes. L'expérience montrerait dans quelle mesure ce résidu tient à des questions purement techniques (par exemple la plus grande difficulté à traiter des syllabes réalisées avec une phonation non modale), et dans quelle mesure il soulève des questions intéressantes pour la modélisation.

2.3 Une nouvelle direction : les dictionnaires en ligne

La réalisation de dictionnaires en ligne, associés aux textes, fait partie du projet de la Collection Pangloss depuis ses débuts. Dans le cadre d'un projet coordonné par G. Jacques[4], trois nouveaux dictionnaires sont en ligne, librement consultables au format HTML ainsi que sous forme de documents PDF (composés en LaTeX). Le dictionnaire japhug-chinois-français comporte près de 8.000 entrées, ce qui en fait un modèle du genre. Une version consultable via smartphone est en cours de réalisation. Ces dictionnaires suivent la norme ISO LMF (Lexical Markup Framework), conçue pour permettre un traitement automatique. L'emploi du format-pivot XML permet des passerelles vers d'autres standards, tels que TEI (Text Encoding Initiative) (Romary 2013). Une lemmatisation systématique des textes permettrait d'accéder à partir du dictionnaire à toutes les occurrences du mot en contexte, et inversement, d'accéder à l'entrée de dictionnaire par un clic sur n'importe quel mot d'un texte. De nombreuses autres pistes sont imaginables pour des collaborations, qui permettraient de réaliser le fort potentiel de ces données.

3 La Collection AuCo

3.1 Présentation

La Collection AuCo est, comme la Collection Pangloss, un projet porté par un centre de recherche (l'Institut International de Recherche MICA, Unité Mixte Internationale CNRS-HUST-Grenoble INP, situé à Hanoi) mais qui a vocation à rendre service à une communauté plus étendue. La Collection regroupe des documents linguistiques sonores du Vietnam et des pays voisins. AuCo est un acronyme pour "Audio Corpora": corpus audio. C'est également une référence à la fée ÂuCơ, qui mit au monde une grande poche d'où sortirent cents œufs qui donnèrent naissance aux Cent Peuples, ancêtres légendaires de la multitude de groupes ethniques de la région. Les points qui composent le logo de la Collection AuCo sont une allusion à ces cent œufs, symbole de la diversité culturelle et linguistique que reflète la collection.

Le but de la Collection AuCo est de rassembler les documents recueillis par les chercheurs au fil de leur activité de recherche, contribuant ainsi à la documentation du patrimoine humain que représentent les langues. La Collection a aussi vocation à encourager et faciliter les travaux de recherche interdisciplinaires associant ingénieurs et linguistes, autour de techniques communes.

La collection accueille des documents de types très divers, et de valeur patrimoniale très inégale : des récits traditionnels aux documents lus, en passant par les dialogues et les enquêtes de vocabulaire ; des collections uniques datant de plusieurs décennies, et concernant des parlers aujourd'hui en voie de disparition, jusqu'au tout-venant des enregistrements de langues nationales (réalisés ponctuellement pour les besoins d'études phonétiques/phonologiques ou d'outils de traitement automatique). Les utilisations nouvelles et créatives des données sont rarement prévisibles; d'où le choix de ne fermer la Collection AuCo à aucun type de données.

[4] http://himalco.huma-num.fr/dictionaries/

3.2 Exemples d'études dont la base empirique est accessible librement

La Collection AuCo, tout comme la Collection Pangloss, offre aux chercheurs la possibilité d'archiver et rendre accessibles les données sur lesquelles reposent leurs travaux. A titre d'exemple, l'étude d'un dialecte de la langue vietnamienne, Phong Nha, a été entreprise en janvier 2014, et les résultats de l'enquête ont été publiés en 2015 (Michaud, Ferlus & Nguyễn 2015). L'enquête reposait sur une liste de vocabulaire, la liste « EFEO-CNRS-SOAS », disponible en ligne (Pain et al. 2014). Les fichiers audio enregistrés lors de l'enquête ont été annotés, en indiquant pour chaque item son numéro dans cette liste ; un script PRAAT (disponible en ligne) est appliqué aux fichiers TextGrid, pour produire des documents XML multilingues (français, anglais, vietnamien, chinois, khmer) synchronisés avec l'enregistrement. Les documents sont en ligne depuis fin 2015. Ainsi, les lecteurs intéressés de découvrir que la spirante /ð/ du vietnamien moyen est préservée dans le dialecte de Phong Nha (alors que dans le delta du Fleuve Rouge elle s'est confondue avec les consonnes /r/ et /ʒ/ ; toutes trois sont actuellement réalisés /z/) peuvent écouter de nombreux exemples, et se faire une opinion au sujet du périmètre de variation allophonique de ce son en fonction de la voyelle et du ton auxquels il se trouve associé.

Le travail de préparation des données a bien sûr demandé un investissement de temps, mais celui-ci a été réduit au minimum. On a fait l'économie du toilettage des fichiers audio : retrancher les portions de silence, d'apartés entre enquêteurs, de raclements de gorge... Les fichiers audio déposés en ligne ne sont pas prévus pour une écoute linéaire, mais pour un accès direct aux mots annotés, auxquels l'interface de consultation donne un accès direct par un bouton « lecture » associé à chacun des items. On pourrait aller jusqu'à argumenter que la mise en ligne des séances des deux journées d'enquête sans aucune retouche présente un avantage : ces documents illustrent le déroulement d'une enquête, et permettent de connaître le contexte de réalisation de chacun des mots. On pourrait par exemple imaginer de mesurer le temps de réponse : entre l'élicitation (le mot en vietnamien standard, fourni oralement) et la réponse du premier locuteur. Cette information pourrait un jour être utilisée dans le cadre d'une étude statistique des pratiques d'élicitation de vocabulaire.

3.3 Les collections de Michel Ferlus : données de plus de quarante parlers

De septembre 2014 à février 2016, dans le cadre de la Collection AuCo a été réalisé un projet de numérisation intitulé « DO-RE-MI-FA : Données des Recherches linguistiques de Michel Ferlus en Asie du sud-est ». Ce projet concerne l'ensemble des enregistrements audio réalisés par Michel Ferlus au cours de son activité comme « linguiste de terrain », de 1963 à 2003 : environ 200 heures d'enregistrements. Michel Ferlus est un spécialiste de la phonétique historique des langues d'Asie du Sud-Est (voir notamment Ferlus 1992, 1998). Ses données inédites proviennent de plus de 40 parlers, dont un grand nombre étaient jusque-là non documentés. Le projet bénéficiait d'une subvention de la Bibliothèque Scientifique Numérique du Ministère de l'Enseignement Supérieur et de la Recherche, dans le cadre de l'opération de numérisation du patrimoine de l'enseignement supérieur et de la recherche. L'objectif est de transformer le fonds de chercheur de Michel Ferlus en un ensemble documentaire dans les règles de l'art, pleinement exploitable, mis à la libre disposition de la communauté des chercheurs ainsi que d'un public plus étendu.

L'enrichissement de ces données par une annotation multilingue s'appuyant sur les notes de Michel Ferlus constitue une entreprise pour le moyen terme, indissociable de la formation de la jeune génération des chercheurs dans ce domaine scientifique. Cette démarche a été engagée dans le cadre du projet, avec la réalisation d'annotations pour la plupart des documents arem, mường, thổ (groupe

vietique de la famille austroasiatique), ainsi que pour une trentaine de documents de langues tai-kadai, par des personnes dont certaines contribueront à prendre la relève du travail de recherche de Michel Ferlus.

Pour le recollement entre enregistrements et transcriptions, le mode opérationnel actuel est le suivant : les manuscrits de Michel Ferlus sont saisis un par un, et leur contenu publié, soit par l'auteur lui-même, soit par un étudiant-chercheur (doctorant) intéressé à reprendre le flambeau pour l'étude d'une langue en particulier, et ayant une certaine familiarité avec la langue. La justification de ce choix est qu'un chercheur engagé dans l'étude de la langue-cible sait tirer le meilleur parti des notes de terrain, et redresser de lui-même les inévitables petites erreurs ou approximations dans la notation. Ce mode de fonctionnement peut paraître extrêmement contraignant : il se peut qu'il faille attendre plusieurs années avant que les données d'une certaine langue trouvent un utilisateur fort d'une bonne connaissance préalable du domaine linguistique concerné. Pour autant, cette perspective n'est nullement utopique : des collègues intéressés se manifestent plus fréquemment que l'équipe du projet ne l'avait initialement espéré.

Au plan technique, le développement et le déploiement d'un outil pour l'affichage synchronisé de manuscrits (en mode image) avec leur annotation (XML) et l'enregistrement audio du texte lu a été engagé dans le cadre du projet DO-RE-MI-FA. Ce logiciel, EASTLing (Easy Annotation & Synchronization Tool for Linguists), comprend un outil d'édition et un outil de lecture (voir : http://moon-light.fr/). Cette innovation illustre l'utilité de collaborations pour le partage de savoir-faire et l'enrichissement de l'interface de consultation des archives, solidaire d'un enrichissement des ressources.

4 Un mot de conclusion

Cette communication a atteint son but si elle est parvenue à mieux faire connaître des phonéticiens la Collection Pangloss et la Collection AuCo, et à suggérer le fort potentiel que présente une association renforcée entre documentation linguistique et recherche. Dans le détail, utilisations et nouveaux développements sont à inventer entre collègues intéressés : projets de recherche, projets applicatifs, ou encore développement d'interfaces « 2.0 » offrant aux utilisateurs le moyen de contribuer aux collections en signalant des corrections, en ajoutant des informations complémentaires, voire en déposant eux-mêmes de nouveaux documents.

Remerciements

Les réalisations présentées ici sont le produit du travail d'un grand nombre de personnes appartenant à divers corps de métier. On a respecté la consigne d'une publication nominale, mais un choix plus cohérent à nos yeux serait de signer du nom de « l'équipe Pangloss », « l'équipe AuCo » et « l'équipe Cocoon » réunies. Nous souhaitons remercier tout particulièrement Anne Behaghel ; Rémy Bonnet ; Céline Buret ; Eric Castelli ; Hằng Đinh ; Michel Ferlus ; Alexandre François ; Julien Heurdier ; Aimée Lahaussois ; Martine Mazaudon ; Boyd Michailovsky ; Minh-Châu Nguyễn ; Việt Sơn Nguyễn ; Thơ Nông ; Frédéric Pain ; Đỗ-Đạt Trần ; et Trí-Dõi Trần. Nous sommes vivement reconnaissants envers les institutions et structures partenaires suivantes : CNRS-InSHS ; Très Grande Infrastructure de Recherche *Humanités Numériques* (Huma-Num) ; CINES ; CC-IN2P3 ; ANR (projets *Empirical Foundations of Linguistics*, ANR-10-LABX-0083, et *HimalCo*, ANR-12-CORP-0006) ; Bibliothèque Scientifique Numérique (projet DO-RE-MI-FA, 2014-2016).

Références

BOTHOREL, André, Péla SIMON, François WIOLAND & Jean-Pierre ZERLING (1986). *Cinéradiographie des voyelles et des consonnes du français*. Strasbourg: Travaux de l'Institut de Phonétique de Strasbourg.

DO, Thê Dung, Thien Huong TRAN & Georges BOULAKIA (1998). Intonation in Vietnamese. In Daniel Hirst & Albert Di Cristo (eds.), *Intonation systems: a survey of twenty languages*, 395–416. Cambridge: Cambridge University Press.

DO, Thi-Ngoc-Diep, Alexis MICHAUD & 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. *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/.

FERLUS, Michel (1992). Essai de phonétique historique du khmer (du milieu du premier millénaire de notre ère à l'époque actuelle). *Mon-Khmer Studies* 21. 57–89.

FERLUS, Michel (1998). Les systèmes de tons dans les langues viet-muong. *Diachronica* 15(1). 1–27.

JACOBSON, Michel, Nicolas LARROUSSE & Marion MASSOL (2015). La question de l'archivage des données de la recherche en SHS (Sciences Humaines et Sociales). *Archives et données de la recherche (ICA/SUV 2014)*. Paris. http://halshs.archives-ouvertes.fr/halshs-01025106.

LADEFOGED, Peter & Ian MADDIESON (1996). *The Sounds of the World's Languages*. Dirigé par M. Kenstowicz, J. Goldsmith, Nick Clements & D. Steriade. Oxford: Blackwell.

LEROY, Christine & Catherine PARIS (1974). Etude articulatoire de quelques sons de l'oubykh d'après film aux rayons X. *Bulletin de la Société de Linguistique de Paris* LXIX(1). 255–286.

MICHAILOVSKY, Boyd, Martine MAZAUDON, Alexis MICHAUD, Séverine GUILLAUME, Alexandre FRANÇOIS & Evangelia ADAMOU (2014). Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation* 8. 119–135.

MICHAUD, Alexis (2002). Conservation des langues et partage des ressources : le rôle des chercheurs dans la mise en place de banques de données. *XXIVe Journées d'Etude de la Parole*, 153–156. Nancy, France.

MICHAUD, Alexis, Michel FERLUS & Minh-Châu NGUYỄN (2015). Strata of standardization: the Phong Nha dialect of Vietnamese (Quảng Bình Province) in historical perspective. *Linguistics of the Tibeto-Burman Area* 38(1). 124–162. doi:10.1075/ltba.38.1.04mic.

MICHAUD, Alexis, Andrew HARDIE, Séverine GUILLAUME & Martine TODA (2012). Combining documentation and research: Ongoing work on an endangered language. In Xiong Deyi, Eric Castelli, Dong Minghui & Pham Thi Ngoc Yen, (eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)*, 169–172. Hanoi, Vietnam: MICA Institute, Hanoi University of Science and Technology.

NIEBUHR, Oliver & Alexis MICHAUD (2015). Speech data acquisition: the underestimated challenge. *KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik* 3. 1–42.

PAIN, Frédéric, Michel FERLUS, Alexis MICHAUD & Thu Hà PHẠM (2014). *EFEO-CNRS-SOAS word list for linguistic fieldwork in Southeast Asia*. Hanoi: International Research Institute MICA. https://halshs.archives-ouvertes.fr/halshs-01068533/.

ROMARY, Laurent (2013). TEI and LMF crosswalks. *arXiv preprint arXiv:1301.2444*.

THIEBERGER, Nick, Anna MARGETTS, Stephen MOREY & Simon MUSGRAVE (2016). Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21. doi:10.1080/07268602.2016.1109428.