Vers des ressources électroniques interconnectées : Lexica, les dictionnaires de la collection Pangloss

Rémy Bonnet, Céline Buret, Alexandre François, Benjamin Galliot, Séverine Guillaume, Guillaume Jacques, Aimée Lahaussois, Boyd Michailovsky, Alexis Michaud

Vers des ressources électroniques interconnectées : Lexica, les dictionnaires de la collection Pangloss

Bonnet, Rémy*, & Buret, Céline[§], & François, Alexandre[§], & Galliot, Benjamin[§],
& Guillaume, Séverine[§], & Jacques, Guillaume*, & Lahaussois, Aimée°,
& Michailovsky, Boyd[§], & Michaud, Alexis[§]

[§]Langues et civilisations à tradition orale (Lacito, CNRS / Université Sorbonne Nouvelle / INALCO)
*Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO, CNRS / EHESS / INALCO)
°Histoire des Théories Linguistiques (HTL, CNRS / Université Paris Diderot / Université Sorbonne Nouvelle)
{bonnet.remy et buret.celine et b.g01lyon et rgyalrongskad et aimeelah et boyd.michailovsky}
@gmail.com, {alexandre.francois et severine.guillaume et alexis.michaud}@cnrs.fr

1 Introduction : vers des ressources électroniques interconnectées

1.1 Données et corpus dans l'exploration de la diversité linguistique

Traditionnellement, la linguistique de terrain vise à la production de grammaires, dictionnaires, et recueils de textes. Ces trois éléments forment ce que l'on appelle la « trilogie boasienne » (Foley, 1999) par référence au travail fondateur de Franz Boas (Boas, 1902 ; Boas & Swanton, 1911). Rien de computationnel dans cette méthode, formulée à une époque où les chercheurs publiaient leurs travaux sous forme imprimée. Mais un siècle plus tard, les technologies numériques permettent une avancée décisive : par l'ajout de la composante multimédia (enregistrements audio et vidéo), la trilogie est devenue *tétralogie* (Musgrave & Thieberger, 2014). Ce tournant a été pris au Lacito dès 1994, par la fondation de la collection Pangloss (Jacobson, Michailovsky, & Lowe, 2001 ; Michailovsky et al., 2014) – archive multimédia en ligne actuellement en pleine expansion.

L'usage des nouvelles technologies va bien plus loin que la simple publication en ligne de travaux autrefois imprimés. Ce qui est désormais crucial, c'est l'établissement de liens dynamiques entre les quatre volets de la tétralogie : demain, dictionnaires et grammaires pourront non seulement être interconnectés, mais aussi liés aux textes qui forment le cœur des données linguistiques, ainsi qu'aux enregistrements audio et vidéo de parole spontanée. Plus que de *fixer* une langue au moyen de l'imprimé, il s'agit désormais de l'offrir à des modes nouveaux de navigation, en exploitant tout le potentiel de corpus en ligne, y compris par des traitements statistiques. Le projet de *tétralogies connectées* et de grammaires électroniques a été formulé clairement (Maxwell, 2012 ; Nordhoff, 2008). La présente communication expose l'état d'avancement de réalisation de dictionnaires en ligne, étape dans l'entreprise qui consiste à porter le projet de *tétralogies connectées* au stade des réalisations pratiques.

1.2 La collection Pangloss et la « linguistique assistée par ordinateur »

La collection Pangloss regroupe un ensemble de ressources multimédia : enregistrements audio ou vidéo effectués sur le terrain par les chercheurs qui étudient des langues dites « rares », transcriptions annotées des contenus des enregistrements, dictionnaires multimédias et, en projet, des grammaires. Plus de 130 langues, des centaines d'heures d'enregistrements, un millier de documents annotés, le tout librement consultable.

Notre travail s'inscrit dans un contexte où chaque ressource peut être directement connectée aux autres ressources. Ainsi, de nombreux enregistrements audio sont synchronisés avec les textes annotés. Les textes transcrits ont vocation à être reliés aux dictionnaires, de façon à proposer pour chaque entrée une concordance de toutes les occurrences, avec la possibilité d'accéder en un clic à l'exemple en contexte, accompagné de l'enregistrement original.

À l'ère du numérique, l'interconnexion des ressources ainsi que les outils d'exploitation automatique facilitent grandement la confrontation des hypothèses avec les données, ce qui encourage la réalisation d'implémentations logicielles des modèles linguistiques.

2 Lexica : présentation des dictionnaires en ligne

Le nom « Lexica » a été adopté pour la dimension lexicographique de la collection Pangloss.

2.1 Pourquoi de nouveaux outils ? La bibliothèque PYLMFLIB

Pour la création de dictionnaires, les outils tels que Toolbox, Fieldworks et LexiquePro, développés par le *Summer Institute of Linguistics*, sont largement utilisés par les linguistes « de terrain », mais ils manquent de flexibilité, en particulier pour l'inclusion des paradigmes de conjugaison des verbes et la conversion automatique d'une orthographe en une autre. Le format utilisé pour les dictionnaires Toolbox (MDF) présente en outre le désavantage d'avoir une structure implicite et ambiguë.

Un travail de développement informatique a donc été réalisé. Une librairie en langage Python 2, PYLMFLIB (https://pypi.python.org/pypi/pylmflib/1.0), a été développée par Céline Buret (2014-2015), et un outil en langage Python 3, doté d'une interface Qt5, par Benjamin Galliot (2016-2017 ; le développement est encore en cours actuellement). Ces outils implémentent en XML la norme lexicographique *Lexical Markup Framework*, LMF (Francopoulo, 2013), conçue comme un format pivot (Romary, 2013), avec des outils de conversion du format MDF vers LaTeX et HTML. Des versions PDF des dictionnaires sont générées à partir de LaTeX.

2.2 Réalisations : les dictionnaires

Trois dictionnaires multimédias *(« talking dictionaries »)* sont actuellement disponibles via l'interface de la collection Pangloss (http://lacito.vjf.cnrs.fr/pangloss/dictionaries/) : japhug, khaling, et limbu. Trois autres dictionnaires (mwotlap, na et teanu) ne comportent pas encore de liens vers des enregistrements à l'heure actuelle : pour ces dictionnaires, le travail d'interconnexion avec les ressources audio et textuelles de la collection Pangloss est en cours.

Le dictionnaire limbu (limbu-népali-anglais), conçu il y a une quinzaine d'années, a été le premier dictionnaire de la collection Pangloss (Michailovsky, 2002, voir également 2011). Il s'agit d'un dictionnaire multimédia pour lequel les exemples des entrées du dictionnaire sont directement reliés à des enregistrements de récits et d'élicitation de vocabulaire, synchronisés eux-mêmes avec leur annotation textuelle et déposés dans la collection Pangloss. Premier dictionnaire « connecté », il est en cours de conversion vers le format LMF.

Le dictionnaire japhug (japhug-chinois-français) comporte plus de 7 000 entrées. C'est le premier dictionnaire de cette langue, d'une grande importance pour l'étude de la famille sino-tibétaine (Jacques, 2016). Il comporte plus de 4 000 fichiers audio, recueillis spécifiquement en vue de la création du dictionnaire. Ces fichiers sont intégrées au fichier PDF ; ils sont également

disponibles dans la version HTML, par le biais d'un hébergement dans NAKALA, un service de la Très Grande Infrastructure de Recherche Huma-Num.

Le dictionnaire khaling (khaling-népali-anglais) recense tous les verbes primaires de cette langue, avec des exemples de phrases, des définitions en népali et en anglais, et des tableaux de conjugaison pour tous les verbes. (Au sujet de la morphologie de cette langue, voir : Jacques, 2015 ; Jacques, Lahaussois, Michailovsky, & Bahadur Rai, 2012.) Une version papier a été publiée localement par la communauté khaling en février 2016 à Kathmandou, en plus de la version PDF intégrant les fichiers audio et la version en ligne.

La librairie est en train d'être pourvue de nouveaux outils (conversion des dictionnaires vers le format Android, interface graphique) pour permettre la production de dictionnaires multimédias dans d'autres langues étudiées par les linguistes de nos laboratoires et d'ailleurs. Une version « LMF » des dictionnaires mwotlap et teanu, et leurs déclinaisons sous forme HTML et PDF, sont en cours de finalisation. (Au sujet de ces langues, voir François, 2003, 2009.)

3 Morphologie et morphotonologie : générateurs de paradigme et projets de modélisation

Les langues que nous étudions présentent des caractéristiques intéressantes sur le plan de la morphologie et de la morpho-phonologie (notamment en ce qui concerne les tons). Les règles morphologiques et morphophonologiques peuvent être implémentées, ce qui permet une confrontation systématique des données avec les règles proposées, d'où des avancées dans l'analyse. Des générateurs de paradigme permettent également d'inclure des paradigmes exhaustifs dans les dictionnaires. Ainsi, pour le dictionnaire khaling, une série de scripts Perl ont été écrits pour générer les paradigmes de conjugaison et convertir l'alphabet phonétique international en alphabet devanagari. Pour aller plus avant dans la modélisation, nous nous orientons vers l'emploi de *transducteurs à états finis* : voir le traitement du yonaguni par Pellard & Yamada (sous presse). Pour traiter les effets à longue distance de la morphophonologie en langue na (Michaud, 2017), nous prévoyons d'utiliser des langages spécialisés dans la programmation pour la linguistique (SLLP : *Specialized Languages for Linguistic Programming*).

4 Conclusion

Les dictionnaires de la collection Pangloss sont librement disponibles, et utilisables en l'état. Mais le rapide panorama présenté ici voulait surtout insister sur les possibilités qui s'ouvrent pour la suite du travail. On aimerait mentionner en conclusion le fait que la libre diffusion en ligne de *données connectées* dans des langues jusqu'ici pas ou peu dotées informatiquement (Berment, 2004) représente un enjeu soci(ét)al évident. Les dictionnaires de la collection Pangloss s'adressent à la fois aux linguistes et aux locuteurs des langues. Un soin particulier a été apporté au dictionnaire khaling, qui promeut une nouvelle orthographe distinguant toutes les oppositions phonologiques de cette langue et répondant aux exigences des locuteurs de la langue, contrairement à l'orthographe précédente, qui souffrait de nombreux défauts techniques. Au-delà de la conservation et de la mise à libre disposition d'un patrimoine culturel inestimable, des « tétralogies » connectées ouvrent de nombreuses possibilités telles que les environnements numériques personnalisés d'apprentissage (Mangeot, Bellynck, Eggers, Loiseau, & Goudin, 2016).

Remerciements

Nous sommes vivement reconnaissants envers : ANR (projets *Fondements Empiriques de la Linguistique*, ANR-10-LABX-0083, et *HimalCo*, ANR-12-CORP-0006) ; CNRS-InSHS ; et Très Grande Infrastructure de Recherche Humanités Numériques (TGIR Huma-Num).

Références bibliographiques

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"* (thèse). Université Joseph Fourier - Grenoble 1.

Boas, F. (1902). *Tsimshian texts*. Washington: Government Printing Office.

Boas, F., & Swanton, J. R. (1911). Siouan (Dakota). In *Handbook of American Indian Languages I* (pp. 875–965). Washington: Government Printing Office, Bureau of American Ethnology, Bulletin 40.

Foley, W. A. (1999). Compte-rendu de Gerrit van Enk & Lourens de Vries, The Korowai of Irian Jaya, Oxford studies in anthropological linguistics, 9, 1997. *Language in Society*, *28*(3), 470–472.

François, A. (2003). *La sémantique du prédicat en mwotlap, Vanuatu*. Louvain: Peeters.

François, A. (2009). The languages of Vanikoro: Three lexicons and one grammar. In *Discovering history through language: papers in honour of Malcolm Ross* (pp. 103–126).

Francopoulo, G. (Ed.). (2013). *LMF: Lexical Markup Framework*. Wiley Online Library.

Jacobson, M., Michailovsky, B., & Lowe, J. B. (2001). Linguistic documents synchronizing sound and text. *Speech Communication, 33 [special issue: "Speech Annotation and Corpus Tools"]*, 79–96.

Jacques, G. (2015). Derivational verbal morphology in Khaling. *Bulletin of Chinese Linguistics*, *8*(1), 78–85.

Jacques, G. (2016). Le sino-tibétain: polysynthétique ou isolant? *Faits de Langues*, *47*(1), 61–74.

Jacques, G., Lahaussois, A., Michailovsky, B., & Bahadur Rai, D. (2012). An overview of Khaling verbal morphology. *Language and Linguistics*, *13*(6), 1095–1170.

Mangeot, M., Bellynck, V., Eggers, E., Loiseau, M., & Goudin, Y. (2016). Exploitation d'une base lexicale dans le cadre de la conception de l'ENPA Innovalangues. In *Enseignement des Langues et TAL* (Vol. 9, pp. 48–64). Paris: Association Francophone de la Communication Parlée.

Maxwell, M. (2012). Electronic grammars and reproducible research. In S. Nordhoff (Ed.), *Electronic Grammaticography* (pp. 207–235). Honolulu: University of Hawaii Press.

Michailovsky, B. (2002). *Limbu-English dictionary of the Mewa Khola dialect, with English-Limbu index*. Kathmandu: Mandala Book Point.

Michailovsky, B. (2011). Limbu. In D. Kouloughli & A. Peyraube (Eds.), *Encyclopédie des sciences du langage, Dictionnaire des langues* (pp. 1064–1074). Paris: Presses Universitaires de France.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., & Adamou, E. (2014). Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation*, *8*, 119–135.

Michaud, A. (2017). *Tone in Yongning Na: lexical tones and morphotonology*. Berlin: Language Science Press.

Musgrave, S., & Thieberger, N. (2014, November). *Rethinking grammatical description: from Heath to hypertext*. Lecture, Research Unit for Indigenous Language, University of Melbourne. Disponible : https://indiglang.arts.unimelb.edu.au/events/rethinking-grammatical-description-from-heath-to-hypertext/

Nordhoff, S. (2008). Electronic reference grammars for typology: challenges and solutions. *Language Documentation and Conservation*, *2*(2), 296–324.

Pellard, T., & Yamada, M. (sous presse). Verb morphology and conjugation classes in Dunan (Yonaguni). In F. Kiefer, J. P. Blevins, & H. Bartos (Eds.), *Morphological paradigms and functions*. Leiden: Brill.

Romary, L. (2013). TEI and LMF crosswalks. *arXiv Preprint arXiv:1301.2444*.