ParCoLaF. Une plateforme de constitution et de diffusion de corpus parallèles pour les langues de France

Dejan Stosic, Myriam Bras

Délégation lénérale à la langue française et aux langues de France - AAP « Langues et numérique 2017 »
Octobre 2018

ParCoLaF

Une plateforme de constitution et de diffusion de corpus parallèles pour les langues de France

Dejan Stosic
Université de Toulouse
CLLE-ERSS UMR5263 UT2J & CNRS
dejan.stosic@univ-tlse2.fr

Myriam Bras
Université de Toulouse
CLLE-ERSS UMR5263 UT2J & CNRS
myriam.bras@univ-tlse2.fr

De nos jours, les dictionnaires sont loin d'être les seuls outils offrant des solutions de traduction d'une langue à l'autre. De nombreuses plateformes numériques permettent en effet d'obtenir en quelques millisecondes la traduction non seulement de mots isolés, comme dans les dictionnaires, mais aussi de séquences, voire de phrases entières. De tels outils d'aide à la traduction, plus ou moins automatisés, sont de plus en plus répandus, surtout pour une petite minorité de langues dominant la toile. Si le français est plutôt bien (ou du moins, de mieux en mieux) loti en la matière lorsqu'il s'agit de le mettre en correspondance avec l'anglais, l'espagnol, l'italien et quelques autres langues bien outillées, sa mise en parallèle avec l'occitan, le breton, le picard ou l'alsacien s'avère beaucoup plus complexe. Ainsi, pour trouver des traductions contextualisées en occitan d'une séquence donnée du français, ou *vice versa*, le recours à la méthode manuelle semble toujours s'imposer, celle-ci consistant soit à fouiller des éditions bilingues de textes, soit à apparier l'original et la traduction. Au-delà de l'inaccessibilité de solutions de traduction à portée de main du grand public pour ces paires de langues, le recueil de grandes quantités de données par des linguistes en vue d'études empiriques comparatives est pratiquement impossible. La constitution d'un corpus parallèle[1] comportant des données du français et des langues de France peut être un moyen de remédiation à ces difficultés.

L'objectif général du projet ParCoLaF est de développer une plateforme pour la constitution, la diffusion et l'interrogation de corpus parallèles comportant des textes en français et en langues régionales de France, avec une focalisation sur l'occitan, qui servira de langue pilote. Le projet vise ainsi la production d'une ressource textuelle électronique multilingue comportant des textes en français et en occitan, alignés au niveau des paragraphes et des phrases avec leurs traductions. La langue occitane fait partie des langues dites « peu dotées ». Quelques projets récents visent à la doter des premières ressources lexicales et textuelles (Bras & Vergez-Couret 2016) ainsi que des outils nécessaires au traitement automatique, en particulier dans le cadre du projet RESTAURE où des chercheurs travaillant sur l'occitan, le picard et l'alsacien mutualisent leurs compétences et leurs efforts pour doter ces trois langues de lexiques et de corpus textuels.

Le projet ParCoLaF résulte des synergies des travaux menés parallèlement au sein de CLLE-ERSS par les membres de deux projets différents, BaTelÒc[2] et ParCoLab[3], le premier se proposant de constituer une base textuelle en langue occitane (3,37 millions de mots actuellement), le second de constituer un corpus parallèle comportant des textes dans trois langues d'Europe (français, anglais, serbe) alignés avec leurs traductions dans les deux autres langues (11 millions de mots actuellement). S'appuyant sur la mutualisation des compétences, savoir-faire, données et ressources propres à chacun des deux projets, le projet ParCoLaF a pour ambition d'offrir une infrastructure à la fois performante et conviviale au service des langues régionales de France en vue de leur mise en valeur mais aussi pour collecter et centraliser des données bilingues et multilingues indispensables autant à la recherche fondamentale en linguistique et en traitement automatique des

[1] Un corpus parallèle peut être défini comme une collection de documents sous forme numérique où chaque unité du texte en langue source est mise en correspondance avec son équivalent en langue cible, au niveau des paragraphes, phrases ou mots.
[2] http://redac.univ-tlse2.fr/bateloc (cf. Bras & Vergez-Couret 2016)
[3] http://parcolab.univ-tlse2.fr (cf.) (cf. Miletic, Stosic & Marjanovic 2018)

langues qu'au développement de différents types d'outils d'aide à la traduction, de dictionnaires électroniques bilingues, et de ressources pour la didactique comparative des langues. La mise à la disposition des chercheurs et du grand public de textes occitans alignés sur des traductions en français, voire en anglais, constitue une innovation majeure pour la langue occitane.

Grâce à un financement dans le cadre de l'AAP « Langues et numérique 2017 », une refonte substantielle de la plateforme ParCoLab a pu être réalisée. Celle-ci peut désormais accueillir jusqu'à 12 langues différentes et offre des fonctionnalités de recherche avancées. L'interface de consultation utilise un site web adaptatif (« responsive web design ») au format HTML5. Ces technologies permettent de consulter la ressource aussi bien sur des ordinateurs que sur des tablettes et smartphones, l'interface s'adaptant dynamiquement et en temps réel au format du support sur lequel elle est consultée.

Parallèlement aux développements informatiques, un travail de recensement et d'intégration de textes disponibles en occitan et en français est mené en collaboration avec les deux partenaires du projet, le *CIRDOC* et *Joliciel*[4]. Au total, une trentaine de textes existant en occitan, en français et/ou anglais ont pu être identifiés[5]. Plusieurs d'entre eux sont en cours d'intégration dans le corpus parallèle, ce qui implique au préalable la négociation des droits d'auteurs, éventuellement la numérisation de documents, la structuration des textes selon les standards actuels en matière de constitution et de diffusion de corpus (format XML, normé TEI5), leur alignement avec les versions disponibles dans d'autres langues et la vérification manuelle de celui-ci. A l'issue de ce processus, les textes alignés au niveau des paragraphes et des phrases sont rendus disponibles dans la base textuelle pour interrogation. La ressource textuelle produite est hébergée par la TGIR Huma-Num, ce qui garantit la pérennisation du patrimoine linguistique et culturelle mis en valeur dans le cadre du projet.

Le projet ParCoLaF ouvre la possibilité d'intégrer plusieurs langues régionales de France dans une plateforme textuelle plurilingue. Il montre que les langues peu dotées en ressources lexicales et textuelles peuvent s'appuyer sur les langues mieux dotées pour exister dans le monde numérique, existence cruciale pour leur survie à l'ère du numérique. Au-delà de la valorisation du patrimoine linguistique et culturel des régions de France dans l'univers numérique grâce à la diffusion des données sous une forme nouvelle (base textuelle alignée), plusieurs types de retombées sont attendus : retombées scientifiques (surtout en linguistique descriptive et comparative et en traitement automatique des langues), pédagogiques (conception d'outils pour la didactique comparative des langues régionales, fabrication automatique d'exercices en ligne, etc.), et applicatives (développement d'outils d'aide à la traduction, de mémoires de traduction pour le français et les langues de France, de dictionnaire en ligne, etc.).

Bibliographie

Bras, Myriam, & Vergez-Couret, Marianne. 2016. BaTelÒc: A text base for the Occitan language. *In* Ferreira, V., Bouda, P. (eds.) *Language Documentation and Conservation in Europe*, Honolulu: University of Hawai'i Press. pp. 133-149.

Miletic, Aleksandra., Stosic, Dejan & Marjanović, Saša. (2017). ParCoLab: A Parallel Corpus for Serbian, French and English. *In* Ekštein K., Matoušek V. (eds), *Text, Speech, and Dialogue. TSD 2017*. Lecture Notes in Computer Science book series, vol. 10415. Springer, Cham, pp. 156-164.

[4] Le *CIRDOC* est le *Centre interrégional de développement de l'occitan* (http://www.locirdoc.fr), *Joliciel informatique* est une entreprise de création de logiciels (http://www.joli-ciel.com).

[5] A titre d'exemple, on peut citer parmi les textes écrits en occitan : R. Lafont, *La gacha a la cistèrna* (Le guetteur à la citerne), B. Manciet, *Lo gojat de noveme* (Le garçon de novembre), M. Rouquette, *Verd paradis* (Vert paradis), traduits du français en occitan : A. de Saint-Exupéry, *Le Petit Prince*, J. Giono, *L'homme qui plantait des arbres*, traduits de l'anglais en occitan : R. Kipling, *The Jungle book* (Le livre de la jungle), A.C. Doyle, *The hound of the Baskerville* (Le chien de Baskerville), R.L. Stevenson, *Treasure Island* (L'Île au trésor).