ORTOLANG[1] : une infrastructure de mutualisation de ressources linguistiques écrites et orales

Jean-Marie Pierrel[1, 2]
(1) Université de Lorraine, ATILF, 44 avenue de la Libération 54063 Nancy Cedex
(2) CNRS, ATILF, 44 avenue de la Libération 54063 Nancy Cedex
Jean-Marie.Pierrel@atilf.fr, contact@ortolang.fr

Résumé. Nous proposons une démonstration de la Plateforme de l'Equipex ORTOLANG (Open Resources and Tools for LANGuage : www.ortolang.fr) en cours de mise en place dans le cadre du programme d'investissements d'avenir (PIA) lancé par le gouvernement français. S'appuyant entre autres sur l'existant des centres de ressources CNRTL (Centre National de Ressources Textuelles et Lexicales : www.cnrtl.fr) et SLDR (Speech and Language Data Repository : http://sldr.org/), cette infrastructure a pour objectif d'assurer la gestion, la mutualisation, la diffusion et la pérennisation de ressources linguistiques de type corpus, dictionnaires, lexiques et outils de traitement de la langue, avec une focalisation particulière sur le français et les langues de France.

Mots-clés : Ortolang, plateforme, mutualisation, corpus, ressources linguistiques

1 Pourquoi une telle infrastructure ?

Une analyse de l'évolution des sciences du langage et du traitement automatique des langues montre que la confrontation avec l'informatique a permis de définir de nouvelles approches. Ainsi au-delà d'une simple linguistique descriptive s'est développée une *linguistique formelle* qui propose des modèles s'appuyant sur une double validation, *explicative* d'un point de vue linguistique, *opératoire* d'un point de vue informatique. Une véritable *linguistique de corpus* permet aussi au linguiste d'aller au-delà de l'accumulation de faits de langue et de confronter ses théories à l'usage effectif de la langue. Ainsi l'informatique est devenue un outil indispensable pour :

- étudier la langue et ses propriétés grâce à l'exploitation de corpus de grande ampleur ;
- structurer et normaliser les connaissances linguistiques (de l'acoustique, à la sémantique) ;
- valoriser et partager les résultats de la recherche grâce à la production de ressources et d'outils informatiques.

Dans ce cadre, les aspects de ressources informatisées (corpus annotés, lexiques et outils de traitement) sont particulièrement importants et stratégiques pour servir de support à la fois :

- aux travaux de recherche pour lesquels la notion de corpus d'étude et de ressources est incontournable ;
- à la diffusion des résultats de ces travaux grâce à leur disponibilité sur la toile.

Un équipement d'excellence de mutualisation de ressources et d'outils pour le traitement informatisé et la valorisation de notre langue s'impose aujourd'hui pour les raisons suivantes :

- Le coût de définition et de production de ressources linguistiques de qualité ou d'outils d'analyse est important. Sans une mutualisation de telles ressources, chaque chercheur se verrait dans l'obligation de tout réinventer !
- L'évaluation de nos productions de recherche (modèles, systèmes de traitement) nécessite la disponibilité de ressources de référence (corpus, lexiques, dictionnaires) accessibles, partagées et clairement identifiables.
- Le partage et la patrimonialisation des connaissances sur les langues de France sont nécessaires afin de faciliter des études sociolinguistiques sur les parlers de France et de les faire bénéficier des apports de la recherche.

2 Principales caractéristiques d'ORTOLANG

Le consortium portant le projet ORTOLANG regroupe des compétences complémentaires en

- sciences du langage à travers l'ATILF, le LPL, MoDyCo et le LLL,
- informatique avec le LORIA et l'INIST, mais aussi en partie l'ATILF et le LPL,
- base de données et accès à de l'information scientifique, à travers l'INIST, et à des ressources linguistiques, à travers les deux centres de ressources que sont le CNRTL et le SLDR.

[1] ORTOLANG bénéficie d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-11-EQPX-0032

Au-delà de la réunion de ces compétences disciplinaires différentes, notre objectif fut aussi de fédérer pour cet équipement de mutualisation de ressources et d'outils sur la langue des partenaires représentant la diversité des approches d'étude de la langue : modélisation linguistique, linguistique expérimentale et/ou appliquée, production et perception du langage, études diachroniques, sociolinguistiques, traitement automatique des langues, écrit, oral.

Cette proposition s'appuie aussi sur une expérience acquise des équipes proposant cet équipement d'excellence et sur une bonne insertion tant nationale qu'internationale :

- acquis des partenaires, centres de ressources et laboratoires, qui alimentent la version initiale de la plateforme avec un ensemble de ressources et d'outils déjà disponibles en leur sein et dont les compétences recouvrent les trois principaux aspects visés : l'oral, l'écrit et la patrimonialisation des parlers de France.
- implication et cohérence avec la TGIR HumaNum.
- implication et cohérence avec l'infrastructure européenne CLARIN.
- cohérence avec les efforts de la DGLFLF et de la BNF sur les aspects patrimonialisation des parlers de France.

La plateforme ORTOLANG est une infrastructure de mutualisation pour la gestion, la pérennisation et la diffusion de corpus et d'outils sur la langue, ces derniers restant bien entendu propriété des déposants (chercheurs ou laboratoires). Nous avons, de plus, prévu des moyens pour aider des laboratoires à finaliser et normaliser leurs ressources.

Quant aux droits d'accès à ces ressources, ils restent donc définis par leurs propriétaires. Toutefois sur ce point ORTOLANG émet des recommandations fortes :

- respect de la charte éthique Big Data, fruit d'un travail collectif réunissant plusieurs acteurs impliqués dans la création, la diffusion et l'utilisation de données,
- liberté d'usage pour la recherche et tant qu'il n'y a pas de valorisation contractuelle,
- moyennant royalties auprès des propriétaires des ressources dès qu'il y a valorisation contractuelle.

C'est dans cet esprit que divers contacts avec des partenaires ayant déposé ou souhaitant déposer leurs ressources sur ORTOLANG ont déjà été mis en œuvre.

3 Objectifs et missions de cette infrastructure

3.1 Identification et préparation des données

Une des difficultés actuelles pour repérer et accéder à des ressources (corpus, dictionnaires, lexiques et outils de traitement) sur notre langue réside dans leur grande dispersion (il n'est pas aisé de savoir quelles ressources sont disponibles et à quels endroits elles sont accessibles) et leur forte disparité, en particulier en termes de codage. Sans compter que nombre de ressources langagières de qualité, développées dans le cadre de projets de recherche ou de thèses, ont été perdues faute d'une gestion rigoureuse de ce patrimoine. C'est pourquoi l'un des premiers objectifs concerne : le catalogage des ressources et outils existants à travers un ensemble de métadonnées normalisées, le contrôle et la validation des ressources et des outils, avec en particulier un accompagnement de leurs auteurs sur les standards, les normes et les recommandations internationales actuelles, et l'enrichissement de ressources et d'outils.

3.2 Pérennisation des ressources

Afin d'assurer la pérennisation des ressources, nous avons mis en œuvre trois types d'actions : la curation des ressources et des outils ; un stockage sécurisé et une maintenance des ressources ; un archivage pérenne, à travers la solution mise en place par la TGIR HumaNum en lien avec le CINES.

3.3 Diffusion

Enfin, pour assurer la nécessaire diffusion et exploitation de ces ressources nous prévoyons une aide et un accompagnement des utilisateurs pour la mise en place des procédures permettant à des utilisateurs de la plateforme d'exploiter ces ressources et outils mutualisés en nous appuyant sur l'expérience des équipes porteuses de l'Equipex et centres de ressources CNRTL et SLDR appelés à terme à se fondre au sein d'ORTOLANG.

4 Démonstration

Au cours de cette démonstration, après une présentation de l'architecture matérielle et logicielle mise en place pour ORTOLANG, nous présenterons la plateforme accessible aujourd'hui dans sa version 0 à l'adresse www.ortolang.fr, ses développements futurs ainsi que les modes d'interactions et de coopérations que nous souhaitons mettre en place avec les producteurs de ressources et d'outils ainsi qu'avec l'ensemble de la communauté utilisatrice potentielle de cette infrastructure.