Pour une meilleure valorisation et mutualisation de ressources linguistiques : quelques acquis de l'Equipex ORTOLANG 1

Jean-Marie Pierrel

Pour une meilleure valorisation et mutualisation de ressources linguistiques : quelques acquis de l'Equipex ORTOLANG[1]

Jean-Marie Pierrel

Université Lorraine & CNRS, ATILF UMR 7118, Nancy France

Ce chapitre est consacré à l'Equipex ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue / *Open Resources and Tools for LANGuage*), mis en place, dans le cadre du programme d'investissement d'avenir (PIA) du gouvernement français, pour permettre une meilleure valorisation et mutualisation de ressources linguistiques (Corpus, Lexiques, Terminologies et Outils de traitement). Après avoir explicité les raisons qui nous ont amenés à structurer ce projet, nous commencerons par rappeler les principales caractéristiques d'ORTOLANG avant de préciser l'originalité et le caractère novateur du projet puis nous présenterons ses principaux services et le flux du travail mis en œuvre dans ORTOLANG pour les phases de dépôt et de publication des ressources.

Nous terminerons en présentant quelques enseignements tirés notre expérience, plus particulièrement en ce qui concerne les processus mis en oeuvre pour aider les utilisateurs à définir et standardiser les métadonnées liées aux ressources déposées sur la plateforme.

1 Pourquoi un Equipex de mutualisation de ressources linguistiques ?

L'usage de plus en plus généralisé de l'informatique dans les études et recherches en sciences humaines et sociales pour permettre d'exploiter au mieux les vastes gisements d'information a fortement contribué au cours des dernières décennies au développement de ce qu'on a coutume d'appeler les humanités numériques. C'est particulièrement vrai en sciences du langage. Une rapide analyse de l'évolution des sciences du langage et du traitement automatique des langues (TAL) au cours des trente dernières années montre en effet que la confrontation avec l'informatique a permis de définir de nouvelles approches. C'est ainsi qu'au-delà d'une simple linguistique descriptive s'est développée une linguistique formelle, couvrant aussi bien les aspects lexicaux que syntaxiques ou sémantiques, qui tend à proposer des modèles s'appuyant sur une double validation, explicative d'un point de vue linguistique, opératoire d'un

[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.

point de vue informatique. C'est elle aussi qui a permis l'émergence d'une véritable linguistique de corpus (Habert et col., 1997) permettant 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. Cette évolution a provoqué une véritable révolution qui fait de l'informatique 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 (acoustiques, phonétiques, morphologiques, lexicales, syntaxiques, sémantiques, etc.) ;
- valoriser, partager et mutualiser les résultats de la recherche sur la langue qui passent le plus souvent par la production de ressources et d'outils informatiques.

Par ailleurs, nous pensons que dans notre société de l'information d'aujourd'hui, seules les langues fortement outillées et modélisées, permettant des traitements automatiques, auront des chances de subsister comme langues véhiculaires de travail et d'échange dans les domaines scientifiques, économiques, industriels et culturels, les autres risquant de se voir réduites à une dimension uniquement vernaculaire. Aujourd'hui, contrairement à ce que quelques esprits chagrins prétendent en affirmant que seul un « anglais international » pourra subsister comme langue véhiculaire et de travail, les jeux sont loin d'être faits. Il est donc important et urgent de doter d'autres langues, dont le français, des outils indispensables à son traitement automatique si nous souhaitons qu'à l'avenir ces langues continuent à jouer un rôle majeur sur le plan intellectuel, économique et sociétal, tant dans le monde industriel que dans celui de la recherche ou de la culture.

Dans ce cadre, les aspects de ressources informatisées (corpus annotés, lexiques, terminologies 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 spécifiquement en linguistique de corpus, en traitement automatique des langues et en didactique des langues ;
- à la diffusion des résultats de ces travaux : l'un des aspects essentiels aujourd'hui est leur informatisation et leur disponibilité sur la Toile sous une forme facilement accessible et exploitable par l'ensemble de la communauté scientifique et industrielle.

Un équipement d'excellence de mutualisation de ressources et d'outils pour le traitement informatisé et la valorisation du français et des langues partenaires s'imposait donc pour les raisons suivantes :

- D'abord, le coût de définition et de production de vastes ressources linguistiques de qualité (corpus, dictionnaires et lexiques), de même que celui de la mise au point d'outils d'analyse (morphologique, morphosyntaxique, lexicale, syntaxique et sémantique) est important. Ce serait un gâchis énorme de vouloir, pour chaque projet de linguistique ou de TAL, redéfinir l'ensemble des ressources dont on a besoin. Sans vouloir plaider pour une rentabilisation maximale de la recherche, il convient de prendre conscience que, sans une mutualisation de telles ressources dans le domaine du langage qui nécessite d'aborder des aspects aussi divers que la phonétique, le lexique, la syntaxe, la sémantique et la pragmatique, chaque équipe de recherche ou chaque chercheur se verrait dans l'obligation de tout réinventer, alors même que nul ne peut être spécialiste de chacun de ces sous-domaines.

- Un second point plaidant pour la mutualisation de ressources concerne l'évaluation de nos productions de recherche (modèles, analyseurs, systèmes de traitement), qui nécessite, pour des besoins de comparaison, la disponibilité de ressources de référence (corpus textuels, lexiques, dictionnaires) accessibles, partagées et clairement identifiables.
- De plus, 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 faire bénéficier ces dernières des apports de la recherche.
- Enfin, en termes de valorisation et de partage de connaissances avec nos concitoyens, une disponibilité accrue, en particulier sur le Web, de nos productions de recherche est indispensable. Outre le fait que cela peut permettre un meilleur partage entre le monde de la recherche et celui de l'entreprise, cela répond aussi à un besoin, de plus en plus grand, de connaissance chez nos concitoyens. Il suffit pour s'en convaincre de voir le nombre de requêtes servies aujourd'hui par le portail lexical du CNRTL[2] (Centre National de Ressources Textuelles et Lexicales) : plus de 600 000 requêtes par jour sur le lexique du français, dont plus de la moitié venant de l'étranger[3] !

Ce sont ces considérations qui nous ont amenés à proposer l'équipement d'excellence ORTOLANG (*Open Resources and Tools for LANGuage* / Outils et Ressources pour un Traitement Optimisé de la LANGue : www.ortolang.fr) de mutualisation de ressources linguistiques.

2 Principales caractéristiques d'Ortolang

2.1 Une ouverture pluridisciplinaire forte

L'Equipex ORTOLANG (Pierrel, 2014) s'appuie sur un consortium constitué de laboratoires et de centres de ressources possédant des compétences complémentaires dans les domaines suivants :

- les sciences du langage à travers quatre unités mixtes de recherche du domaine : l'ATILF (Analyse et Traitement Informatique de la Langue Française : www.atilf.fr), le LPL (Laboratoire Parole et Langage : www.lpl-aix.fr) , MoDyCo (Modèle, Dynamiques, Corpus : www.modyco.fr) et le LLL (Laboratoire Ligérien de Linguistique : www.lll.cnrs.fr) ;
- l'informatique avec le LORIA (Laboratoire LOrrain de Recherche en Informatique et ses Applications : www.loria.fr) et l'INIST (INstitut de l'Information Scientifique et Technique : www.inist.fr), mais aussi en partie l'ATILF et le LPL, deux laboratoires SHS d'interface avec l'informatique ;
- la maîtrise des bases de données et de l'accès à de l'information scientifique, à travers l'INIST, ainsi que des ressources linguistiques, au travers de deux centres de ressources créés par le CNRS en 2006 : le CNRTL (Centre National de Ressources Textuelles et Lexicales, porté par l'ATILF : www.cnrtl.fr) (Pierrel & Petitjean, 2007) et le SLDR (Speech & Language Data Repository, porté par le LPL : www.sldr.org).

[2] http://www.cnrtl.fr/portail/

[3] http://www.cnrtl.fr/aide/stat/

Au-delà de la réunion de ces compétences disciplinaires différentes, notre objectif est 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, sociolinguistique, traitement automatique des langues (écrit, oral et multimodal).

S'appuyant sur les acquis des partenaires, centres de ressources (CNRTL et SLDR) et laboratoires, qui offrent un ensemble de ressources et d'outils disponibles et dont les compétences recouvrent les principaux aspects visés : l'oral, l'écrit, le multimodal et la patrimonialisation des parlers de France, ORTOLANG s'intègre de façon cohérente dans le paysage national et international :

- implication et cohérence avec la TGIR HumaNum[4] ;
- cohérence avec l'infrastructure européenne CLARIN[5] au sein de laquelle nous avons travaillé lors de la phase préliminaire et dont ORTOLANG est appelé à devenir un nœud de son réseau de centres CLARIN (Wittenburg and al., 2010) ;
- cohérence, enfin, avec les efforts menés par la DGLFLF (Délégation Générale à la Langue Fançaise et aux Langues de France) et la BNF (Bibliothèque Nationale de France) sur les aspects patrimonialisation des parlers de France.

2.2 Un équipement gérant des ressources pour l'ensemble de la communauté scientifique

La plateforme ORTOLANG est donc une infrastructure de mutualisation pour la gestion, la pérennisation et la diffusion de ressources et d'outils sur la langue qui, bien entendu, restent propriété des déposants (chercheurs ou laboratoires). Les droits d'accès à ces ressources sont donc définis par leurs propriétaires. Toutefois, sur ce point, ORTOLANG émet des recommandations fortes (cf. la charte d'Ortolang[6]) :

- respect de la charte éthique *Big Data*[7], 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 tant qu'il n'y a pas d'usage commercial ;
- négociation préalable avec les propriétaires des ressources, dès qu'il y a souhait de valorisation commerciale.

2.3 Des objectifs ambitieux

Les principaux objectifs de l'Equipex ORTOLANG, tels que nous les avions définis dès le départ, sont doubles :

a) Servir de support aux travaux de recherche pour lesquels la notion de corpus est aujourd'hui incontournable spécifiquement en linguistique et en traitement automatique du langage.
b) Œuvrer pour la valorisation des résultats de recherche (corpus, lexiques, dictionnaires et outils de traitement). Comme nous l'avons déjà indiqué, un des

[4] www.huma-num.fr

[5] https://www.clarin.eu/

[6] https://www.ortolang.fr/information/policy

[7] http://wiki.ethique-big-data.org.

aspects essentiels aujourd'hui est leur informatisation et leur disponibilité sur la Toile sous une forme facilement accessible et exploitable par l'ensemble de la communauté scientifique et industrielle.

Ces objectifs nécessitent d'assurer le partage et la pérennisation des ressources produites et cela d'autant plus qu'elles jouent un rôle central dans de nombreux domaines de recherche en linguistique et en TAL. Elles sont en effet nécessaires pour l'émergence et l'apprentissage de modèles, que ce soit dans le cadre des approches stochastiques, ou des approches symboliques. Elles sont aussi indispensables pour valider et évaluer les théories et outils résultats de travaux de recherche et dans ce cadre, il est indispensable de disposer de ressources de référence ou « étalons » supports de ces évaluations.

Notons aussi que ces objectifs répondent à un souci d'efficience et d'économie de la recherche. Constituer des ressources linguistiques entraine des coûts non négligeables, il convient donc d'éviter de refaire deux fois la même chose. À titre d'exemple, le coût de la constitution de corpus annotés en linguistique de corpus et en TAL est généralement évalué à 0,10 € le mot pour des corps écrits annotés et à 0,50 € le mot pour des corpus oraux ou multimodaux, alors que, par exemple, les approches stochastiques requièrent des corpus de plusieurs millions de mots !

2.4 Une architecture matérielle et logicielle solide et sécurisée

Afin de permettre un service 24h/24, 7j/7, 365j/an avec un taux de disponibilité de haut niveau, nous avons choisi d'implanter l'architecture matérielle d' ORTOLANG à l'INIST. Elle repose sur des moyens spécifiquement acquis par le projet (serveurs, système d'exploitation, disques durs, robotique de sauvegarde) et des moyens INIST partagés (réseau, pare-feu, hyperviseur, réseau de stockage et de sauvegarde, serveur de pilotage des sauvegardes, salles machines).

L'architecture matérielle spécifique[8] mise en place dans le cadre du projet s'appuie sur :

- Un cluster de serveurs composés de 6 serveurs : 3 R620 – 48 coeurs – 768 GB de mémoire vive (RAM) et 3 serveurs R630 – 60 coeurs – 1152 GB de mémoire vive (RAM).
- Un système de stockage configuré en utilisant des mécanismes de redondance et correction d'erreur Raid 6 offrant 165 téraoctets utiles de disques. Le cluster de serveurs est connecté au système de stockage par un réseau haut débit *Fiber Channel* à double attachement. La connectique réseau et le *firewall* apportés par l'infrastructure INIST sont eux-mêmes sécurisés et redondants, assurant une haute disponibilité de la plateforme.
- Un système de sauvegarde s'appuyant sur une librairie Quantum avec 2 lecteurs LTO6 et 50 slots de 300 To.

Nous avons choisi d'utiliser des technologies de virtualisation pour avoir le maximum de souplesse et exploiter au maximum les ressources physiques (puissance CPU, capacité de la mémoire centrale RAM, pool de stockage). Ainsi les serveurs physiques hébergent des machines virtuelles qui peuvent être déplacées d'un serveur à l'autre pour assurer des maintenances programmées et la continuité en cas de défaillance matérielle.

[8] https://dev.ortolang.fr/doc.infrastructure.html

Quant à l'architecture logicielle, elle s'appuie sur un centre de diffusion compatible avec les recommandations de l'infrastructure européenne CLARIN[9] (European Research Infrastructure for Language Resources and Technology) pour ses centres de ressources (Wittenburg *et coll.*, 2010) sur lequel se greffe directement le site Web www.ortolang.fr permettant aux utilisateurs de naviguer dans les ressources ou de sélectionner des ressources via des requêtes sur les métadonnées. L'ensemble du code développé est disponible en open source.

Cette architecture logicielle a été mise en place pour supporter des contraintes de qualité de service (disponibilité maximale) et de gestion des ressources permettant d'obtenir le DSA[10] (*Data Seal of Approval*). Le cœur du système, entrepôt OAI-PMH peu visible des utilisateurs, est donc un dépôt fiable de données intégrant les fonctionnalités suivantes:

- identification de chaque ressource par un identifiant pérenne (ou *Handle*) ;
- preuve d'intégrité de la donnée associée à un identifiant pérenne fournie sous forme d'un ensemble de contrôles liée à l'identifiant pérenne ;
- gestion de versions : toute modification d'une donnée doit conduire à une nouvelle version. Cette gestion des versions s'effectue à travers une relation dédiée dans les métadonnées ;
- authentification des utilisateurs à travers un mécanisme de signature unique (*Single Sign On*) utilisant la fédération Education-Recherche de Renater[11] lors de la consultation de données à accès restreint ;
- implémentation de la notion de déposant, en dédiant un élément à cet effet dans les métadonnées, un déposant pouvant être un individu, un projet, un laboratoire ou une institution.

ORTOLANG représente aussi un investissement en logiciel de dépôt, de gestion et de diffusion de données *open source* en plus d'être (en premier) une source de données. Aujourd'hui nous disposons d'une plateforme logicielle à fort potentiel de valorisation comme solution au moins équivalente, voire supérieure en terme de couverture fonctionnelle, à d'autres plateformes existantes (telle par exemple DSpace[12]) et potentiellement réutilisable par d'autres entités souhaitant disposer d'une plateforme de dépôt d'objets numériques. Lors du dernier colloque annuel CLARIN, nous avons pu noter un intérêt fort de plusieurs communautés de recherche européennes pour disposer d'une telle plateforme.

2.5 ORTOLANG : un Equipex parfaitement intégré dans le paysage national des sciences du langage

2.5.1 Un service spécialisé de la TGIR Huma-Num pour la langue

Nous travaillons en étroite collaboration avec Huma-Num et notamment pour l'exposition et la valorisation des données déposées sur ORTOLANG via Isidore.

[9] https://www.clarin.eu/

[10] http://datasealofapproval.org

[11] https://services.renater.fr/federation/index

[12] https://www.dspace.com/fr/fra/home.cfm

Aujourd'hui, suite à l'accord avec Huma-Num, ORTOLANG est un service spécialisé pour les langues, complémentaire de l'offre généraliste proposée par Huma-Num. La TGIR a inclus l'Equipex comme membre de son réseau de partenaires au côté du CLEO et des MSH[13]. Ainsi, lorsque le TGIR Huma-Num reçoit des demandes relevant des sciences du langage, ces dernières sont systématiquement transmises à ORTOLANG.

Par ailleurs, suite à la collaboration étroite avec les consortiums de linguistiques, il a été décidé pour le nouveau consortium CORLI « Corpus, Langues et Interactions, qui a pris la suite des consortiums « Corpus Ecrits » et « IRCOM », d'adjoindre ORTOLANG comme partenaire pour les actions de dépôt, conservation et diffusion de corpus langagiers et de finalisation de corpus, en particulier au travers d'appels à propositions lancés conjointement.

2.5.2 Des liaisons étroites avec les fédérations CNRS en Sciences du langage

ORTOLANG, dès sa validation, a choisi de mettre en place des liaisons étroites avec les deux fédérations en Sciences du langage :

- la fédération ILF (Institut de linguistique Française) : ORTOLANG, focalisé au départ sur la langue française, est complémentaire de la fédération ILF et doit servir de plateforme d'accueil pour le projet « corpus de référence du français »[14]
- la fédération TUL (Typologie et Universaux linguistiques), ORTOLANG a en effet fait le choix de s'ouvrir vers l'accueil et la gestion des corpus des chercheurs français en linguistique, quelle que soit la langue étudiée.

Notons par ailleurs que, spécialisé sur le partage, la diffusion et la pérennisation de ressources linguistiques, ORTOLANG accueille des ressources langagières issues d'autres communautés scientifiques que les sciences du langage, en particulier l'histoire (cf. par exemple le corpus AMPLOR[15]) et la littérature. Nous sommes d'ailleurs pleinement ouverts vers une accentuation de ces efforts d'accueil de ressources langagières issues d'autres composantes des SHS.

3 Originalité et caractère novateur du projet

Projet validé dans le cadre du programme d'investissements d'avenir (PIA), ORTOLANG a pour mission d'offrir un réservoir de ressources et d'outils clairement disponibles et documentés permettant de remplir un double objectif de partage de connaissance et de mutualisation d'acquis. Cette infrastructure doit permettre à la communauté de franchir un pas décisif, aujourd'hui encore à peine ébauché. Il s'agit non seulement de regrouper des contenus et une variété de données ou d'outils disponibles, mais aussi et surtout d'assurer la diffusion de standards clairs, internationalement reconnus, pour les données comme pour les métadonnées afin de pouvoir les rendre accessibles et permettre le partage, la réutilisation et la complémentation des dites ressources. L'intérêt d'une telle infrastructure peut en fait s'analyser selon plusieurs points de vue complémentaires détaillés ci-dessous.

[13] http://www.huma-num.fr/presentation/réseau

[14] http://www.ilf.cnrs.fr/spip.php?rubrique95

[15] https://www.ortolang.fr/market/corpora/amplor

3.1 Intérêt pour la communauté de recherche en linguistique

Depuis une dizaine d'années, le paysage de la recherche en linguistique a largement évolué grâce à l'apparition d'importants corpus de langue aisément disponibles sur Internet. Si l'existence d'une linguistique de corpus n'est pas nouvelle (Laks, 2008), cette évolution de l'accès aux données dynamise de manière très importante le domaine, permet de démontrer l'importance, du point de vue fondamental, de la notion de variation, et autorise de grandes avancées dans la modélisation des théories exemplaristes ou dites "basées sur l'usage" (Tomasello, 2000 ; Barlow and Kemmer, 2000).

Si avant les années 2000, le paradigme générativiste dominait et conduisait à voir les théories et les modèles linguistiques comme fondamentalement sous-déterminés par les données factuelles, ce n'est plus le cas aujourd'hui. Comme le note Newmeyer (2003), ce sont d'abord les travaux psycholinguistiques d'observation longitudinale, et spécialement ceux menés sur les acquisitions précoces, qui ont ébranlé le paradigme cognitiviste chomskyen en documentant une hétérogénéité et une variabilité intrinsèque très importantes et peu compatibles avec l'innéisme de la grammaire universelle. Ces travaux ont récemment rencontré les problématiques de la linguistique variationniste conduites indépendamment depuis plusieurs décennies. La confrontation avec les analyses du changement linguistique en temps réel a par ailleurs souligné l'importance des dynamiques qui structurent, forment et déforment les systèmes linguistiques dans le temps. Enfin, le développement des travaux contrastifs et typologiques a conduit à relativiser la portée des grandes hypothèses universalistes au profit d'une description plus fine et plus précise des données observées. Dans chacun des domaines et des sous-domaines des sciences du langage, la notion d'usages ou de pratiques attestées a ainsi été remise au premier plan, induisant un rapport nouveau aux modélisations explicatives et aux formalisations (Barlow & Kemmer, 2000).

Ces théories, tels les *pattern grammars* (Hunston & Francis, 2000) qui puisent leurs racines dans les travaux de Sinclair (1991) en linguistique de corpus, sont basées sur la notion de constructions, qui sont des associations entre forme et fonction. Les constructions peuvent être extrêmement variées, allant de formes figées (un mot, une holophrase, une expression idiomatique) à des structures plus générales (par exemple la structure transitive sujet-verbe-objet), et en passant par de nombreux intermédiaires plus ou moins généralisés (par exemple la construction "c'est X" où "X" peut prendre n'importe quelle forme ; ou la construction "X aime Vinf" où "X" et "Vinf" sont mutuellement contraints). Les constructions peuvent se combiner pour produire des formes langagières de tout niveau de complexité. De telles théories permettent de modéliser la variété à tous les niveaux, de l'interlocuteur à l'intralocuteur. Elles font évoluer le système de catégorisation mis en place sur les exemplaires connus en élargissant sa base empirique, en modifiant le poids fréquentiel d'une série d'exemplaires, en favorisant la formation d'une construction plus générale que celles qui étaient disponibles sous la forme d'exemplaires auparavant.

L'apport de la linguistique de corpus à la compréhension des phénomènes langagiers est donc devenu fondamental. Grâce à l'augmentation de la variété et de la taille des corpus, il est aujourd'hui devenu possible de démontrer les faits langagiers à l'aide d'exemples attestés en grand nombre et de tester les propositions de la linguistique et de la psycholinguistique sur des données de productions véritables, mais pour cela, un grand nombre de corpus contrôlés, bien décrits et variés, est nécessaire.

3.2 Intérêt d'une telle proposition pour la communauté de TAL

La multiplication des corpus offre également de nouvelles ouvertures hors du champ de la linguistique et de la psycholinguistique, en matière de simulation et de traitement automatique du langage naturel aussi bien écrit qu'oral ou multimodal. En effet, la majorité des traitements automatiques réalisés aujourd'hui sur le langage naturel s'appuie sur des approches d'analyse de grandes masses de données et exploite des modèles construits sur ces mêmes corpus. Cette nécessité d'avoir accès à de grandes bases de données se retrouve également dans les méthodes d'évaluation des modèles ainsi conçus. Ceux-ci requièrent des statistiques suffisantes pour garantir la validité des performances des modèles automatiques ainsi que leur robustesse aux diverses sources de variabilité du langage rencontrées en conditions réelles d'application. La comparaison de différents modèles théoriques et la participation aux campagnes d'évaluation qui tendent à se multiplier dans le domaine du TAL requièrent également de grandes quantités de données. Elles participent sur le long terme à formaliser ce domaine de recherche et contribuent significativement à sa progression, comme l'illustre par exemple l'évolution du champ d'application de la transcription automatique de la parole au cours de ces dernières décennies (Haton & col., 2006) .

La mise à disposition pérenne de grands corpus normalisés et enrichis comme le propose ORTOLANG constitue ainsi un progrès très important pour la communauté de recherche en TAL et un accélérateur certain pour les recherches menées dans ces domaines. Ainsi, pour la reconnaissance automatique de la parole, domaine de recherche dont la progression est structurée et rythmée par les campagnes d'évaluations sur des corpus payants dédiées successivement, par exemple, aux informations radiophoniques (ESTER[16]) et aux émissions de télévision (ÉTAPE[17]), l'ambition unanimement affichée consiste à diversifier les styles de paroles et à ouvrir les évaluations aux enregistrements de réunions (*Meetings*) et aux conversations spontanées (*Switchboard*), comme cela a déjà été réalisé aux États-Unis par le NIST[18] dans le cadre des campagnes RT[19]. ORTOLANG a pour ambition de permettre la mise en place et la distribution de telles données d'étude.

Un autre exemple en TAL concerne les recherches en analyse syntaxique automatique, qui souffrent, particulièrement en France, du manque de corpus dédiés aux différents genres du français notamment oral. La récente campagne d'évaluation PASSAGE[20] des analyseurs syntaxiques illustre les besoins de la communauté en grandes masses de données annotées, comme l'a démontré dans le reste de l'Europe la succession des campagnes CoNLL[21].

Les volets constitution, enrichissement et diffusion de corpus constituent donc, là aussi, une base de travail unique et de grande valeur pour la communauté française du domaine.

---

[16] http://www.afcp-parole.org/camp_eval_systemes_transcription/

[17] http://www.afcp-parole.org/etape.html

[18] NIST : National Institute of Standards and Technology, http://www.nist.gov/

[19] RT : Rich Transcription Evaluation Project, https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

[20] http://atoll.inria.fr/passage/eval2.fr.html

[21] http://ifarm.nl/signll/conll/

3.3 Intérêt du point de vue culturel et pédagogique

La diffusion de données de langage, contrôlées et validées, est également fondamentale du point de vue culturel et pédagogique.

Du point de vue culturel, pour la diffusion du patrimoine de la langue française, des langues de France et des langues en contact avec le français, l'existence de ressources fiables et finement décrites est fondamentale. En particulier, depuis 1911, année de l'inauguration par Ferdinand Brunot des *Archives de la parole* en France, qu'il a créées avec l'aide d'Émile Pathé, la conservation des enregistrements sonores et des documents écrits qui leur sont liés est une préoccupation qui repose sur une relation entre les chercheurs et les institutions de conservation. Si cette question est aujourd'hui intégralement traitée, dans le cas de documents édités, par le biais du dépôt légal des archives sonores dont la BNF a la responsabilité, il n'en est pas de même pour les corpus électroniques produits et exploités par les chercheurs dont le dépôt reste souvent difficile, voire impossible, pour des raisons techniques et juridiques, d'autant qu'ils ne correspondent que rarement aux produits commerciaux qui ont retenu l'attention du législateur (musiques, dialogues de film, etc.).

Sur un plan technique, les besoins pour les opérations de catalogage sont la mise en place de descripteurs à intégrer dans une ontologie qui reste à construire et une indication déclarative des codages utilisés. Le catalogage doit prendre en compte les liens qui existent entre des données primaires audio ou vidéos et l'incrémentation des transcriptions et annotations qui leur sont liées dès lors qu'il s'agit de corpus ouverts, évolutifs ou dynamiques. Sur un plan juridique, la prise en compte des conditions de conservation et d'exploitation permet de résoudre les problèmes liés à la protection de la vie privée (données personnelles, droit moral) et à la gestion des droits patrimoniaux et de propriété intellectuelle.

Du point de vue de l'enseignement et de l'apprentissage des langues, l'existence de données bien décrites, comprenant des métadonnées détaillées (y compris par exemple la description du contexte pragmatique de production du corpus), peut servir de source précieuse pour les supports audiovisuels ainsi que pour les supports d'enseignement à distance à une époque où la référence à des « documents authentiques » a enfin supplanté les « exemples construits » ou « exemples d'école » (Duda & Tyne, 2012). La disponibilité de telles données est donc nécessaire pour l'amélioration des supports de cours, par exemple en apprentissage du français langue seconde ou langue étrangère.

3.4 Intérêt du point de vue des partenariats public privé

Les applications industrielles de la linguistique, notamment en matière d'accès à l'information, de structuration de connaissances, majoritairement sous formes langagières, de didactique des langues et de dialogue homme-machine, sont dépendantes de la qualité et de la taille des corpus d'apprentissage et de référence dont elles disposent. Ces recherches ont un impact d'un point de vue économique, à travers les entreprises de logiciels ou de communication homme-machine, et toutes celles qui créent des produits utilisant le support du langage humain (oral comme écrit, souvent associés) et qui exploitent ou ont besoin de données de qualité et de grande taille sur lesquelles développer leurs produits. Or la plupart des entreprises du domaine, start-ups et PME, ne peuvent se permettre, compte tenu des coûts d'investissement à prévoir, d'élaborer des ressources linguistiques à large couverture. Une telle infrastructure devrait permettre aux partenaires industriels de tester des ressources, lors des phases

de recherche et de développement de prototypes. Une rémunération par royalties des producteurs de ces ressources intervenant ensuite dès que l'utilisation de ces dernières conduit à une exploitation commerciale.

Ainsi une telle infrastructure doit permettre aussi d'aider le tissu industriel français à développer ses outils de traitement de la langue sans nécessiter un ticket financier d'entrée souvent incompatible avec les charges de nos start-ups ou PME.

4 Les services offerts

Les services d' ORTOLANG se déclinent en trois aspects complémentaires : identification et préparation des données, enrichissement de ressources et d'outils, pérennisation des ressources.

4.1 Identification et préparation des ressources

L'une des difficultés actuelles pour repérer et accéder à des ressources (corpus, dictionnaires, lexiques, terminologies et outils de traitement) réside tout à la fois dans leur grande dispersion et leur forte disparité, en particulier en termes de codage. De plus, au cours des vingt dernières années, 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 :

- La finalisation et standardisation de ressources et d'outils existants en vue de leur mutualisation. Cette action est menée en étroite coopération avec les consortiums Ecrit et IRCOM et maintenant CORLI de la TGIR Huma-Num. Afin de créer un tel mouvement de mutualisation largement ouvert vers des équipes externes au consortium, nous avons mis en place, dans le cadre du projet ANR support de notre Equipex, des financements au travers d'appels à projets communs avec les consortiums linguistiques pour soutenir les nécessaires travaux de normalisation de ressources que les équipes externes au consortium souhaitent déposer sur la plateforme ORTOLANG.
- Le contrôle et la validation des ressources et des outils, avec en particulier un accompagnement des auteurs de ressources sur les standards, les normes et les recommandations internationales actuelles tels XML, TEI[22], LMF[23] et SYNAF[24].
- L'enrichissement de ressources et d'outils. Cette action s'appuie sur les équipes porteuses d'ORTOLANG et concerne, entre autres, le développement d'un concordancier travaillant sur de gros volumes et utilisable sur tout corpus de langue écrite, l'enrichissement d'un lexique morphosyntaxique du français, l'amélioration de la couverture temporelle d'un lemmatiseur du français et sa mise à disposition sous forme de Web Service, le développement d'outils de segmentation de phrases multilingues, le développement d'outils d'aide à la transcription de corpus oraux, le développement de *plugins* assurant l'interopérabilité entre les différents outils d'édition et d'annotation, le

[22] www.tei-c.org/

[23] http://www.lexicalmarkupframework.org/

[24] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37329

développement d'une grammaire couvrante du français et enfin la normalisation de divers corpus parmi lesquels COLAJE[25], l'Est Républicain[26], ESLO[27], PFC[28], TCOF[29].

4.2 Pérennisation des ressources

Afin d'assurer la pérennisation des ressources, nous avons mis en oeuvre trois types d'actions :

- la curation des ressources et des outils, et en particulier leur normalisation dans des standards internationalement reconnus ;
- l'hébergement des ressources numériques liées à la langue et à son traitement (corpus, dictionnaires, lexiques, terminologies et outils de traitement) permettant une organisation des objets dans des collections, un enrichissement des métadonnées et un catalogue des objets disponibles ;
- un stockage sécurisé et une maintenance des ressources incluant en particulier une identification unique des objets, un contrôle de l'accès aux objets, une gestion de l'historique des états des objets au travers de la notion de versions d'une ressource ;
- un archivage pérenne, à travers la solution mise en place par la TGIR Huma-Num en lien avec le CINES.

4.3 Diffusion et partage des ressources

Le troisième service offert par ORTOLANG concerne la diffusion et le partage de ressources. Nous proposons une aide et un accompagnement des utilisateurs pour la mise en place des procédures leur permettant d'exploiter les ressources et les outils mutualisés en nous appuyant sur les expériences précédentes des centres de ressources CNRTL et SLDR.

Notons enfin qu'outre les efforts déployés pour faciliter le travail de standardisation et normalisation des ressources avant publication sur lequel nous allons revenir ci-après, ORTOLANG collecte aussi des statistiques et assure la diffusion de notifications d'usage auprès des déposants.

5 Le flux du travail mis en œuvre dans ORTOLANG pour les phases de dépôt et de publication des ressources

Un effort particulier a été mené pour offrir une interface et des espaces de travail proposant aux déposants une procédure souple et la plus conviviale possible pour permettre à des non-informaticiens de facilement déposer et valoriser leurs ressources. L'hébergement, le stockage et l'archivage des ressources ne sont en effet qu'une partie du processus, la phase de dépôt et de mise en forme de la ressource est capitale. Pour ce

[25] https://www.ortolang.fr/market/corpora/colaje

[26] https://www.ortolang.fr/market/corpora/est_republicain

[27] https://www.ortolang.fr/market/corpora/eslo1

[28] https://www.ortolang.fr/market/corpora/pfc

[29] https://www.ortolang.fr/market/corpora/tcof

faire nous avons développé un flux de travail ou *workflow*[30] qui se décompose en 5 étapes (cf. figure 1).

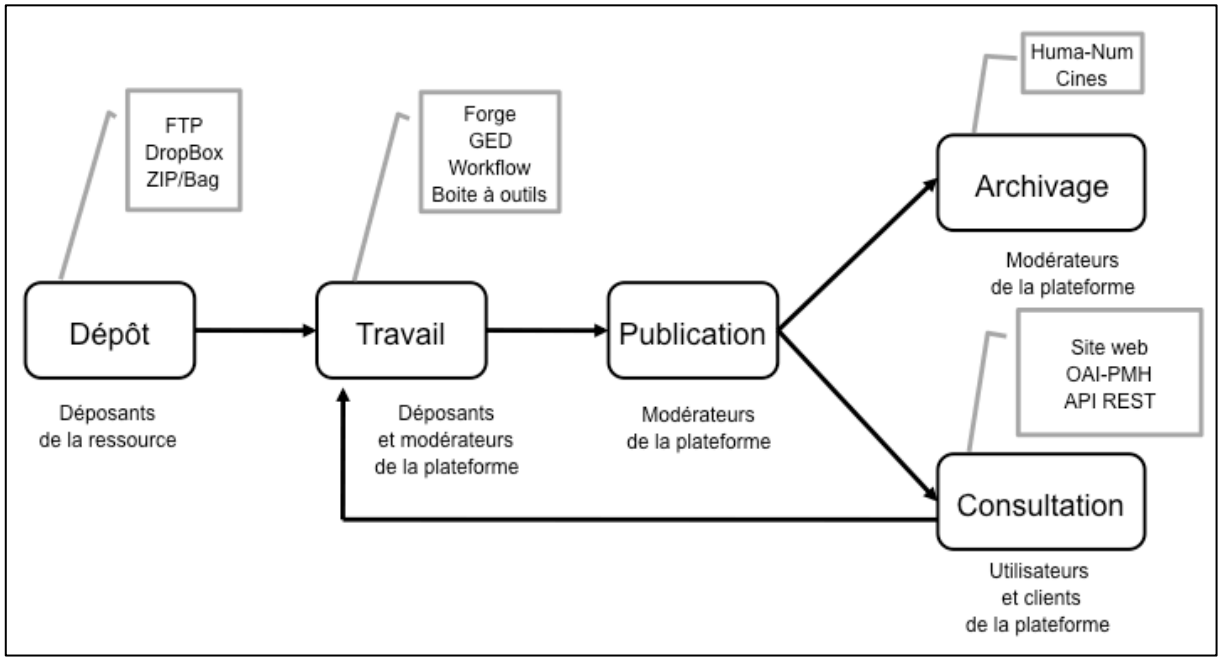

Figure 1 : Flux de travail de dépôt d'une ressource dans ORTOLANG

1. **Dépôt** : Les utilisateurs qui souhaitent déposer une nouvelle ressource doivent s'identifier sur la plateforme en se créant un compte et ouvrir un espace de travail dans lequel ils déposent l'ensemble de leurs données et de leurs métadonnées (fichiers). Pour cette phase initiale de dépôt plusieurs techniques sont mises à leur disposition : dépôt web de fichiers ou d'archive zip au travers d'une sorte de *Dropbox* et liaison FTP entre autres. Hélas, ces données brutes, fruit d'un travail de recherche, ne sont pas toujours prêtes à être publiées en l'état, car elles ne répondent pas forcément aux contraintes de publication ou d'archivage. Néanmoins, aussitôt déposées, ces ressources sont sécurisées par la plateforme au travers de l'utilisation de supports fiables (redondance) et de sauvegardes incrémentales quotidiennes sur bande.
2. **Travail au sein de l'espace de travail sécurisé**. Les déposants peuvent alors assurer un travail de mise en forme en collaboration étroite avec les administrateurs de la plateforme, c'est l'étape de travail. Cette étape, particulièrement importante, concerne en particulier la standardisation des données et la définition des métadonnées sur laquelle nous allons revenir dans le paragraphe suivant. Durant cette phase de travail, l'accès aux données est contrôlé : elles ne sont visibles que par les membres de l'espace de travail et les administrateurs de la plateforme. De plus les producteurs de ressources peuvent bénéficier du support de trois centres de compétences : l'un plus orienté vers des ressources pour l'écrit (ATILF/CNRTL), le second vers des ressources pour l'oral (SLDR & Modyco), et le troisième vers des ressources multimodales (SLDR & Modyco). Le producteur dispose de plus au sein de son espace de travail de divers outils en ligne lui permettant de

[30] https://dev.ortolang.fr/doc.workflow.html

- déclarer les membres de cet espace de travail qui doivent s'être préalablement ouvert un compte sur la plateforme ;
- déposer de nouveaux contenus et spécifier pour chaque dossier ou fichier déposé leur visibilité externe. Nous avons choisi de limiter à quatre les choix d'accessibilité et de visibilité des ressources : (i) à tous (ii) aux utilisateurs connectés et donc s'étant préalablement déclarés sur la plateforme (iii) aux seuls membres de l'ESR (Enseignement Supérieur et Recherche) (iv) restreint aux membres de l'espace de travail (comme précisé dans la charte d'ORTOLANG ce dernier niveau très restrictif doit être justifié soit par des raisons d'exploitation de la ressource et de retombées des travaux, soit par des raisons juridiques) ;
- enrichir son travail (conversion de format, alignement, annotation, etc.) et en particulier définir les métadonnées correspondantes ;
- et, une fois les métadonnées définies, obtenir une prévisualisation de sa ressource.

3. **Publication**. Lorsque les données sont prêtes et les métadonnées définies, le producteur peut soumettre une requête de publication qui est prise en compte par l'équipe d'ORTOLANG pour vérifier entre autres la standardisation du codage et la cohérence entre données et métadonnées. En effet dès la publication, il nous faut garantir la pérennité des données : elles ne changeront plus du moins pour cette version. Durant cette phase le déposant peut suivre l'état de sa demande et, en collaboration avec les équipes d'ORTOLANG au travers de fils de discussion attachés à chaque espace de travail, aboutir à une version stable de sa ressource.
4. **Archivage**. Les données publiées peuvent être soumises pour un archivage à long terme via la solution proposée par Huma-Num en lien avec le CINES. L'enrichissement automatique des données pendant les phases antérieures a permis de disposer de données « propres » et le format d'archivage a été vérifié. Cet archivage n'est pas systématique, il se fait après validation conjointe d'ORTOLANG et d'Huma-Num. Il convient en effet de bien distinguer le stockage sécurisé, assuré par ORTOLANG, de l'archivage pérenne qui a un coût non négligeable et ne se justifie que principalement dans trois cas (i) lorsque la ressource est unique et ne pourrait plus être redéfinie, c'est en particulier le cas pour des enquêtes sociolinguistiques telles ESLO1[31] sur les parler dans la région d'Orléans dans les années 1960 (ii) pour des ressources liées à des langues ou parlers en voie de disparition (iii) lorsque la reconstruction de la ressource demanderait un effet financier et humain supérieur à celui de son archivage pérenne.
5. **Consultation et réutilisation**. Une fois publiées, la consultation des ressources peut se faire de plusieurs manières dont une via l'interface Web proposée sur la plateforme qui présente toutes les ressources hébergées, organisées par catégories et décrites par une fiche détaillée. Une navigation dans le contenu des ressources est également disponible en ligne, ainsi que des possibilités de téléchargement si l'utilisateur dispose des droits spécifiés par la licence attachée à la ressource. Les données publiées peuvent par ailleurs être référencées dans un nouvel espace de travail.

[31] https://www.ortolang.fr/market/corpora/eslo1

6 Quelques enseignements tirés de notre expérience

6.1 Nécessité d'une standardisation du codage des ressources

Comme nous l'avons indiqué ci-dessus, il est particulièrement important d'aller vers une standardisation du codage des ressources. Sans une telle standardisation, il n'est en effet pas possible ni de réutiliser facilement ces ressources ni de les outiller véritablement.

Dans le cadre d'ORTOLANG, nous privilégions un codage en XML/TEI, pour l'écrit, et Transcriber, wav, clan, MP3, pour l'oral, afin d'offrir des outils facilement applicables à cet ensemble de ressources. Un des premiers outils incontournables dans ce cadre est un *vieweur* ou système de visualisation des ressources de type corpus permettant aux utilisateurs d'avoir un aperçu convivial du contenu de la ressource.

Prenons à titre d'exemple le cas de transcription de manuscrits avec codage en TEI de la source et de la transcription : il est alors possible tout à la fois de proposer (i) une visualisation au travers d'une feuille de style standard ou spécifique proposée par le déposant (ii) une visualisation de la source TEI, souvent utile uniquement pour les spécialistes (iii) une visualisation au travers d'un outil générique tel que CETEIcean[32] inspirée par Boilerplate[33] (Cayless & Viglianti 2016) qui est une bibliothèque JavaScript permettant l'affichage d'un fichier TEI dans un navigateur Web.

C'est cette dernière technique que nous avons par exemple utilisée pour les ressources multimodales, combinant texte et image et codées en TEI, comme par exemple *Corpus 14*[34] et *AMPLOR*[35]. Lorsqu'on accède à ces ressources à l'aide du bouton « Parcourir » sur la plateforme et que l'on sélectionne un fichier XML/TEI, la première visualisation correspond à une visualisation simple au travers d'une feuille de style standard, cas de *Corpus 14,* ou spécifique et donc plus riche proposée par l'équipe déposante, comme c'est le cas pour *AMPLOR.* Lors de cette première visualisation deux boutons spécifiques apparaissent : « Voir le code source » qui permet de visualiser le codage XML/TEI et « Prévisualisation CETEIcean » qui propose une visualisation plus claire du texte et des images associées.

6.2 Gestion et standardisation des métadonnées

Les métadonnées, ensemble structuré de données créées pour fournir des informations sur des ressources électroniques, sont aujourd'hui nécessaires pour la mise en place de différentes fonctions dont, entre autres :

a) le signalement et l'accès aux ressources au travers d'informations sur le contenu de la ressource pour en faciliter la découverte, la localisation, l'accès : ce sont elles en particulier qui assurent l'amélioration de la pertinence et de l'exhaustivité des recherches, le tri et le filtrage des résultats ;
b) la gestion des ressources décrites (suivi du cycle de vie : création, modification, archivage) ;

[32] http://teic.github.io/CETEIcean/

[33] http://dcl.ils.indiana.edu/teibp/

[34] https://www.ortolang.fr/market/corpora/corpus14

[35] https://www.ortolang.fr/market/corpora/amplor

c) la description et le suivi des droits et conditions d'utilisation associés à la ressource ;
d) la traçabilité de la ressource : historique des mises à jour, des versions successives, autres formats disponibles, sources.

L'analyse des usages des ressources linguistiques nous ont conduits à mener une réflexion à la fois sur le type de métadonnées à gérer, leur standardisation de codage et sur les processus permettant aux déposants de définir de façon la plus efficace et conviviale possible les métadonnées minimales requises pour un dépôt.

Divers schémas de codage de métadonnées étaient préexistants à la mise en place d'ORTOLANG, le plus important étant sans aucun doute celui respectant la norme Dublin Core pour le signalement des ressources. L'objectif du Dublin Core est en effet de fournir un socle commun d'éléments descriptifs pour améliorer le signalement et la recherche de ressources au-delà des diverses communautés et des nombreux formats descriptifs propres à chaque spécialité, tout en restant suffisamment structuré. Il prévoit 15 éléments tous facultatifs et tous répétables, qui portent sur la description :

- du contenu : *Titre* (nom donné à la ressource), *Sujet* (thème du contenu de la ressource), *Description* (présentation du contenu de la ressource), *Source* (référence à une ressource dont la ressource décrite est dérivée), *Langue* (langue du contenu de la ressource), *Relation* (référence à une ressource apparentée), *Couverture* (Périmètre ou domaine d'application du contenu de la ressource) ;
- de la propriété intellectuelle : *Créateur* (entité principalement responsable de la création du contenu de la ressource), *Contributeur* (responsable de contributions au contenu de la ressource), *Éditeur* (entité responsable de la mise à disposition ou diffusion de la ressource), *Droits* (informations sur les droits associés à la ressource) ;
- de l'instanciation : *Date* (date de création ou de mise à disposition de la ressource), *Type* (nature ou genre du contenu de la ressource), *Format* (physique ou numérique de la ressource), *Identifiant* (référence univoque à la ressource).

La norme Dublin Core est indépendante des formats d'encodage et de stockage de l'information. La DCMI[36] (Dublin Core Metadata Initiative) propose néanmoins des recommandations et bonnes pratiques pour permettre l'emploi uniforme de cette norme. Que ce soit en plein texte, dans les pages HTML ou XHTML, notamment pour un meilleur signalement à travers l'emploi du Dublin Core dans les balises *meta* et *link*, dans des documents XML ou aussi dans des assertions en RDF (*Resource Description Framework)*, un modèle de graphe destiné à décrire de façon formelle des ressources et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions.

Ces métadonnées peuvent être ensuite codées et normalisées suivant différents schémas. L'un des schémas le plus répandu correspond à l'OAI-PHM. (Open Archives Initiative's Protocol for Metadata Harvesting) ou protocole OAI mis en place pour faciliter l'échange de données entre des fournisseurs de données et un fournisseur de service (un portail thématique ou local désirant rassembler des données par exemple). Ce protocole d'échange permet de créer, d'alimenter et de tenir à jour, par des procédures automatisées, des réservoirs d'enregistrements qui signalent, décrivent et

[36] http://dublincore.org/

rendent accessibles des documents, sans les dupliquer ni modifier leur localisation d'origine. Grâce au protocole OAI, un fournisseur de données a la possibilité d'offrir une visibilité accrue à ses documents. Réciproquement, un fournisseur de service peut réaliser une base de données ou un portail documentaire dans son domaine de spécialité ou sur un thème quelconque, en collectant les données descriptives de ressources et documents de tous types, accessibles sur l'Internet dans des entrepôts OAI.

De plus, dans différents domaines des extensions du Dublin Core ont été proposées. Ainsi, dans le domaine de la linguistique et du TAL, OLAC[37] propose cinq extensions : *Type de discours* (roman, récit, jeu de langage, etc.), *Identification de la langue* (au travers de son code ISO: fr, en, etc.), *Champ linguistique* (sociolinguistique, phonétique,etc.), *Type de données linguistiques* (lexique, corpus, description de la langue), *Rôles des participants* (annotateur, auteur, locuteur, etc.). La mise en œuvre XML des métadonnées OLAC suit les directives pour la mise en œuvre de Dublin Core en XML[38].

Ainsi divers formats de métadonnées pour les ressources et outils linguistiques coexistent : OAI-PMH, TEI-headers, CMDI, RDF, auxquels il convient d'ajouter d'autres schémas de métadonnées métier plus spécifiques en fonction du type de ressources et de leurs exploitations possibles. C'est d'ailleurs ce qui a conduit CLARIN a lancé l'infrastructure de composants métadonnées (CMDI). Il fournit un cadre pour décrire et réutiliser des plans de métadonnées. Les blocs de construction de description (« composants », qui incluent des définitions de champs) peuvent être regroupés en un format de description prédéfini (ou schéma). Les deux sont stockés et partagés avec d'autres utilisateurs dans le « Registre des composants » pour promouvoir leur réutilisation. Chaque enregistrement de métadonnées est ensuite exprimé sous forme de fichier XML, incluant un lien vers le schéma sur lequel il est basé.

L'une des difficultés principales pour nous a donc été de déterminer quel type de schéma de métadonnées retenir, d'autant qu'il n'est pas évident pour un utilisateur linguiste de dominer l'ensemble des schémas de codage existants. Si bien entendu, nous acceptons et gérons les différents schémas de métadonnées que nous fournissent nos déposants (OAI_DC, OLAC, TEI, RDF), nous avons choisi de définir des métadonnées spécifiques à ORTOLANG, compatibles avec les trois principaux schémas dominant dans la communauté linguistique OAI-PMH, OLAC et CMDI.

Conscients de la difficulté que peut avoir un utilisateur linguiste pour coder de telles métadonnées, en particulier la normalisation des valeurs des divers champs de ces métadonnées, nous avons choisi, dans le processus de publication, d'offrir un outil web interactif lui permettant de définir ces métadonnées sans se soucier de leurs codages spécifiques. Dans un second temps, à partir de ces saisies, nous générons automatiquement, d'une part, un format interne qui nous permet de proposer différents filtres permettant de faire une recherche simple ou avancée dans l'ensemble des ressources que nous gérons et, d'autre part, des projections dans les formats OAI-PHM et OLAC. Ce sont effet ces métadonnées OAI-PMH qui sont ensuite butinées par des services externes, dont Isidore[39], le moteur de recherche unifié des sciences humaines et

[37] http://www.language-archives.org/

[38] http://www.dublincore.org/documents/dc-xml-guidelines/

[39] https://www.rechercheisidore.fr/

sociales mis en place par Huma-Num et moissonnant une grande quantité de liens émanant d'ORTOLANG et de nombreuses bibliothèques numériques.

Lors du dépôt d'une ressource sur ORTOLANG et avant de pouvoir demander la publication de cette ressource, le déposant doit donc introduire les métadonnées correspondantes à l'aide de formulaires web proposés par ORTOLANG. Pour ce faire nous nous appuyions sur des listes déroulantes de recherche dans les référentiels d'autorité nationaux ou internationaux sur lesquels il s'appuie ou, en cas d'absence de référentiels, dans des listes d'autorité spécifiques que nous gérons sur la plateforme pour aider l'utilisateur à remplir les divers champs demandés et obtenir ainsi un codage normalisé de ses métadonnées, tout en laissant la possibilité de soumettre une nouvelle valeur au cas où la valeur recherchée n'apparaitrait pas dans le référentiel correspondant.

Les diverses rubriques des métadonnées ORTOLANG sont :

1. Métadonnées générales : *Nom de la ressource* correspondant au nom de l'espace de travail ; *Catégorie* (Corpus, Lexique, Terminologie, Outil) ; *Titre* ; *Descriptif* ; *Lien* vers la documentation.
2. Personnes impliquées : *Laboratoire producteur* ; *Soutien institutionnel* ; *Personnes contributrices* avec identification et rôle (auteur, concepteur, développeur, locuteur, annotateur, etc.). Concernant l'identification des personnes définies dans leurs profils ORTOLANG, nous leur proposons de renseigner leurs identifiants externes pérennes tels que leurs identifiants IdHal[40], ORCID[41], VIAF[42], IdREF[43], ou ceux de réseaux sociaux professionnels (Linkedln[44], Viadeo[45]).
3. Champs spécifiques suivant la catégorie de la ressource
   - Pour les corpus : *Type* de corpus (Ecrit, oral, multimodal) ; *Type de langue* (monolingue, multilingue, comparable, parallèle) ; *Langues du corpus* ; *Langues étudiées* ; *Genre* (journalistique, littéraire, écrits scientifiques, écrits professionnels, nouveaux modes de communication, parole naturelle, et.) ; *Niveau d'annotation* (transcription, annotation phonétique, conversationnelle, mimo-gestuelle, etc.) ; *Format* ; *Encodage* de caractère ; *Taille* du corpus.
   - Lexique : *type* d'entrées (formes fléchies, lemmes, etc.) ; *Type de langue* (monolingue, bilingue, multilingue) ; *Langues des entrées* ; *Nombre d'entrées* ; *Type de description* (genre, variante féminine, nombre, etc.) ; *Langue de description* ; *Format* (CSV, LMF, MySQL, TEI).
   - Terminologie : *Type* de la ressource (ontologie, taxinomie, liste de termes, thésaurus) ; *Type de structure* (polyhiérarchie, réseau de relations, arborescence stricte, liste à plat) ; *Champs de description* terminologique (contexte, relations hiérarchiques, ontologiques ou associées, forme préférentielle, synonymes, domaine, définition) ; *Contexte* linguistique (mono ou multilingue) ; *Langues des*

[40] https://hal.archives-ouvertes.fr/page/mon-idhal

[41] https://orcid.org/

[42] https://viaf.org/

[43] https://www.idref.fr/autorites/autorites.html

[44] https://fr.linkedin.com/

[45] http://fr.viadeo.com/fr/

entrées ; *Modèles* et *Formats* (ANSI ISOxxxx, OWL, Skos, format spécifique) ; *Origine* de la ressource ; etc.
- Outils : *Système d'exploitation* ; *Langage de programmation* ; *Fonctionnalités* (alignement texte parole, concordancier, indexation, analyse flexionnelle, etc.) ; *Format d'entrée* ; *Format de sortie* ; *Encodage* des caractères ; *Langues traitées* ; *Langues de navigation* ; *Type de support*.

4. Informations complémentaires : *Publication* de référence ; *Aperçu* ; *Mots clés* ; *Site Web* ; *Licence*, *Conditions spécifiques* et *Copyright* ; *Visibilité* attachée à chaque fichier. Nous avons défini quatre types de visibilités des ressources : pour tous, pour les utilisateurs connectés, pour les membres de l'ESR (Enseignement Supérieur et Recherche), pour les seuls membres de l'espace de travail. Pour gérer cette visibilité, nous avons mis en place une procédure d'authentification des membres de l'ERS qui s'appuie sur la fédération d'identité Renater[46].

6.3 Intérêt d'un accompagnement et d'une valorisation les producteurs de ressources

Un troisième enseignement que nous avons tiré de notre expérience ORTOLANG concerne l'intérêt d'un accompagnement et de la valorisation des producteurs de ressources. Concernant le premier point, nous nous appuyons sur trois centres de compétences thématiques, Oral (SLDR & Modyco), Multimodal (SLDR & Modyco) et Ecrit (ATILF/CNRTL). Ainsi, dès la première phase de dépôt, les déposants peuvent avoir des interactions directes et personnelles avec les membres de ces centres de compétences grâce à la mise en place de fils de discussion au sein de chaque espace de travail.

Par ailleurs, nous avons mis en place diverses statistiques d'usage et des informations de suivi régulier sont transmises aux déposants.

Enfin pour mieux valoriser les producteurs de ressources, nous leur offrons la possibilité, au travers de leur profil, de fournir leurs identifiants externes, tel l'IdHal qui permet de présenter automatiquement, en plus de la liste des ressources auxquelles ils ont contribué et qui sont déposées dans ORTOLANG, leurs listes de publications déposées sous HAL. De même pour la valorisation des laboratoires producteurs, nous offrons leur liste permettant, grâce à un simple clic sur leur identifiant dans la liste des producteurs de ressources sur ORTOLANG[47], d'extraire la liste des ressources auxquelles ils ont contribué (cf. https://www.ortolang.fr/producers/atilf).

Nous espérons que cette visibilité et valorisation des déposants et producteurs de ressources dans ORTOLANG pourra contribuer, à terme, à la meilleure prise en compte des efforts de production et mutualisation de ressources des individus et des laboratoires dans les procédures d'évaluation auxquelles ils sont soumis ?

7 Conclusion

La plateforme ORTOLANG est opérationnelle depuis 2016 et accessible via le site www.ortolang.fr. Elle permet de rechercher une ou des ressources grâce à une recherche par facettes s'appuyant sur les métadonnées définies et permettant des

[46] https://services.renater.fr/federation/index

[47] https://www.ortolang.fr/producers

sélections suivant, entre autres, les droits d'utilisation (libres ou sous droits), les langues, les types de ressources (corpus écrits, corpus oraux ou multimodaux, lexiques, terminologies, outils de traitement), les formats, le type d'annotation ou les producteurs des ressources. La plateforme ORTOLANG permet aussi de suivre l'évolution du projet et d'accéder à un ensemble de documentation sur les ressources gérées.

Au final, la réussite d'un tel projet repose bien entendu sur les services et ressources offerts à la communauté, mais aussi et surtout sur l'appropriation par la communauté scientifique de cet outil de mutualisation de ressources linguistiques écrites et orales. Aujourd'hui, ORTOLANG est un Equipex au service de l'ensemble de la communauté sciences du langage. Bien que sa mise en service sous sa forme complètement opérationnelle soit récente (mi-2016), la plateforme réunit déjà des ressources très diversifiées, allant bien au-delà de celles réalisées par ses seuls partenaires. Fin janvier 2019, Ortolang regroupe 402 ressources dont 287 ressources publiées (213 corpus, 16 lexiques, 22 terminologies, 32 outils et 4 projets intégrés) et 70 ressources en construction ou en cours de finalisation ce qui représente 7,1 To de données et plus de 379 000 fichiers. Par ailleurs, étant donné que certaines ressources sont soumises à des restrictions d'accès, en particulier à l'ESR, il est à noter que, fin janvier 2019, 1499 utilisateurs s'étaient créé un compte sur la plateforme Ortolang. De plus, au cours des derniers mois nous avons été contactés par plusieurs laboratoires et divers projets ANR qui souhaitent héberger au sein d'ORTOLANG l'ensemble de leurs ressources et outils. Pour les laboratoires, en plus des partenaires initiaux du projet (ATILF, LLL, LORIA, LPL, MoDyCO) cela concerne deux UMR : le laboratoire BCL (pour héberger leur plateforme Hyperbase), le laboratoire SFL (pour héberger un site Web sur la langue des signes).

Nous pensons que ces données chiffrées sont autant d'indices de réussite de notre Equipex et du service qu'il rend déjà à la communauté scientifique.

Remerciements

Nous tenons à remercier l'ensemble des équipes de l'ATILF, de l'INIST, du LLL, du LPL, du LORIA et de MoDyCo, partenaires d'Ortolang et plus particulièrement l'équipe de développement informatique (Etienne Petitjean, Jerôme Blanchard, Cyril Pestel, Frédéric Pierre) qui ont permis la réalisation de ce projet. Nos remerciements vont aussi au programme d'investissement d'avenir (PIA) du gouvernement français et à l'ANR qui nous ont octroyé les moyens financiers indispensables à la réalisation du projet.

Bibliographie

BARLOW Michael & KEMMER Suzanne. *Usage based models of language*. Chicago : University of Chicago Press, 2000

CAYLESS Hugh & VIGLIANTI Raffaele « CETEIcean, a JavaScript library for isomorphic TEI to HTML transformation », TEI Conference and Members' Meeting 2016, 26th to 30th September, Vienna, Austria,

DUDA Richard & TYNE Henry. Authenticity and autonomy in language learning. *Bulletin Suisse de Linguistique Appliquée*, vol. 92, pp. 86-106, 2012.

HABERT Benoit, NAZARENKO Adeline & SALEM André. *Les linguistiques de corpus*. Paris : Armand Colin., 1997

HATON Jean-Paul, CERISARA Christophe, FOHR Dominique, LAPRIE, Yves & SMAÏLI Kamel *Reconnaissance automatique de la parole - Du signal à son interprétation*. Paris : Dunod, 365 p., 2006

HUNSTON, Susan & FRANCIS Gill. "Pattern Grammar A corpus-driven approach to the lexical grammar of English", *Studies in Corpus Linguistics*, John Benjamins Publishing Company, 288 pp., 2000

LAKS Bernard. "Pour une phonologie de corpus." *Journal of French Language Studies*, vol. 18, n° 1. pp. 3-32, 2008.

NEWMEYER Frederick J. « Grammar is grammar and usage is usage », *Language*, vol. 79. pp. 682-707, 2003

PIERREL Jean-Marie. « Ortolang : Une infrastructure de mutualisation de ressources linguistiques écrites et orales », *Cahiers de l'Acedle Recherches en didactique des langues et cultures*, volume 11, numéro 1, pp. 169-190, 2014

PIERREL Jean-Marie & PETITJEAN Etienne. « Le CNRTL, Centre National de Ressources Textuelles et Lexicales, un outil de mutualisation de ressources linguistiques », *Actes de TALN 2007*, Vol 2, Toulouse, IRIT Press, 12-15 juin 2007, p. 327- 330

SINCLAIR John (1991) *Corpus, Concordance, Collocation*. Oxford University Press.

TOMASELLO Michael. "First steps toward a usage-based theory of language acquisition." *Cognitive Linguistics*, vol. 11, n° 1/2. pp. 61-82, 2000

WITTENBURG Peter, BEL Nuria, BORIN Lars, BUDIN Gerhard, CALZOLARI Nicoletta, HAJICOVA Eva, KOSKENNIEMI Kimmo, LEMNITZER Lothar, MAEGAARD Bente, PIASECKI Maciej, PIERREL, Jean-Marie, PIPERIDIS Stelios, SKADINA, Inguna, TUFIS Dan, VAN VEENENDAAL, Remco, VÁRADI Tamas & WYNNE Martin « Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure ». *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010). Valetta, Malte, 17-23 mai 2010. http://www.lrec-conf.org/proceedings/lrec2010/index.html.