UNE INFRASTRUCTURE DE PÉRENNISATION, DE MUTUALISATION ET DE VALORISATION DE RESSOURCES LINGUISTIQUES : L'EQUIPEX ORTOLANG[1]

JEAN-MARIE PIERREL

Université de Lorraine, CNRS, ATILF
jean-marie.pierrel@atilf.fr

Mots-clés : langage, ressources, mutualisation, préservation à long terme, infrastructure.
Key-words: language, resources, mutualization, long-term preservation, infrastructure.

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 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, 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.

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

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

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

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 leur traitement automatique si nous souhaitons qu'à l'avenir ces langues continuent à jouer un rôle majeur sur le plan intellectuel, économique et sociétal.

Dans ce cadre, les aspects de ressources informatisées (corpus, lexiques, terminologies et outils de traitement) sont particulièrement importants et stratégiques pour servir de support d'une part 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 et, d'autre part, 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. Un équipement de mutualisation de ressources et d'outils pour le traitement 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 de vouloir, pour chaque projet de linguistique ou de TAL, redéfinir l'ensemble des ressources dont on a besoin. Il convient en effet de prendre conscience que, sans une mutualisation de telles ressources dans le domaine du langage, chaque équipe de recherche ou chaque chercheur se verrait dans l'obligation de tout réinventer.

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 accessibles, partagées et clairement identifiables.

De plus, le partage et la patrimonialisation des connaissances sur les langues sont nécessaires afin de faciliter des études sociolinguistiques et de faire bénéficier ces dernières des apports de la recherche.

Enfin, en termes de valorisation et de partage de connaissances, une disponibilité 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] : plus de 600 000 requêtes par jour sur le lexique du français, dont plus de la moitié venant de l'étranger !

[2] Centre National de Ressources Textuelles et Lexicales : http://www.cnrtl.fr/portail/

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

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.

1. PRINCIPALES CARACTÉRISTIQUES D'ORTOLANG

1.1. Une ouverture pluridisciplinaire forte

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 de quatre unités mixtes de recherche du domaine : l'ATILF (Analyse et Traitement Informatique de la Langue Française), le LPL (Laboratoire Parole et Langage), MoDyCo (Modèle, Dynamiques, Corpus) et le LLL (Laboratoire Ligérien de Linguistique) ;

L'informatique avec le LORIA (Laboratoire LOrrain de Recherche en Informatique et ses Applications) et l'INIST (INstitut de l'Information Scientifique et Technique), 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 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) (Pierrel et Petitjean 2007) et le SLDR (Speech & Language Data Repository, porté par le LPL).

Au-delà de la réunion de ces compétences disciplinaires différentes, notre objectif est aussi de fédérer pour cet équipement 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 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 au travers de sa cohérence avec la TGIR Huma-Num[3], et l'infrastructure européenne CLARIN[4] dont ORTOLANG est appelé à devenir un nœud de son réseau de centres (Wittenburg and al., 2010).

[3] www.huma-num.fr
[4] https://www.clarin.eu/

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

1.2. Un équipement au service de l'ensemble de la communauté scientifique

ORTOLANG est 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 aux ressources sont donc définis par leurs propriétaires. Toutefois ORTOLANG émet des recommandations (cf. la charte d'ORTOLANG[5]) : le respect de la charte éthique *Big Data*[6], fruit d'un travail collectif réunissant plusieurs acteurs impliqués dans la création, la diffusion et l'utilisation de données ; la liberté d'usage pour la recherche tant qu'il n'y a pas d'usage commercial ; la nécessaire négociation préalable avec les propriétaires des ressources, dès qu'il y a souhait de valorisation commerciale.

1.3. Des objectifs ambitieux

Les principaux objectifs de l'Equipex ORTOLANG, tels que nous les avions définis dès le départ, sont doubles : (i) 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. (ii) Œuvrer pour la valorisation des résultats de recherche (corpus, lexiques, dictionnaires et outils de traitement). Comme nous l'avons déjà indiqué, 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.

Notons aussi que ces objectifs répondent à un souci d'efficience et d'économie de la recherche. Constituer des ressources linguistiques entraine en effet des coûts non négligeables, il convient donc d'éviter de refaire deux fois la même chose !

1.4. 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). Cette architecture s'appuie sur un cluster de 6 serveurs : 3 R620 – 48 cœurs, 768 GB de mémoire vive (RAM) – et 3 serveurs R630 – 60 cœurs, 1152 GB de mémoire vive (RAM) –, un système de stockage utilisant des mécanismes de redondance et correction d'erreurs offrant 165 téraoctets utiles, un système de sauvegarde s'appuyant sur une librairie Quantum avec 2 lecteurs LTO6 et 50 slots de 300 To. Nous avons de plus 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

[5] https://www.ortolang.fr/information/policy
[6] http://wiki.ethique-big-data.org.

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

peuvent être déplacées d'un serveur à l'autre pour assurer des maintenances programmées et la continuité en cas de défaillance matérielle d'un des composants.

Quant à l'architecture logicielle, elle s'appuie sur un centre de diffusion compatible avec les recommandations de l'infrastructure européenne CLARIN sur lequel se greffe directement le site Web permettant aux utilisateurs de naviguer dans les ressources ou de sélectionner des ressources via des requêtes sur les métadonnées.

Cette architecture logicielle a été mise en place pour supporter des contraintes de qualité de service (disponibilité maximale) et de gestion des ressources. 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 ; authentification des utilisateurs à travers un mécanisme de signature unique (*Single Sign On*) lors de la consultation de données à accès restreint ; implémentation de la notion de déposant (individu, projet, laboratoire ou institution), en dédiant un élément à cet effet dans les métadonnées.

2. ORIGINALITÉ ET CARACTÈRE NOVATEUR DU PROJET

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. Il s'agit non seulement de regrouper des contenus et une variété de données ou d'outils disponibles, mais aussi d'assurer la diffusion de standards, internationalement reconnus, pour les données comme pour les métadonnées afin de pouvoir les rendre accessibles et d'en permettre le partage, la réutilisation et la complémentation. L'intérêt d'une telle infrastructure peut en fait s'analyser selon plusieurs points de vue complémentaires détaillés ci-dessous.

2.1 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" (Barlow and Kemmer 2000).

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

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

la taille des corpus, il est devenu possible de démontrer les faits langagiers à l'aide d'exemples attestés en grand nombre et de tester les propositions de la linguistique ou de la psycholinguistique sur des données véritables, mais pour cela, un grand nombre de corpus contrôlés, bien décrits et variés, est nécessaire.

2.2. Intérêt d'une telle proposition pour la communauté de TAL

La multiplication des corpus offre également de nouvelles ouvertures en matière de simulation et de traitement automatique du langage écrit, 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 corpus. Cette nécessité d'accès à de grandes bases de données se retrouve également dans les méthodes d'évaluation des modèles. Ceux-ci requièrent des statistiques suffisantes pour garantir la validité de leurs performances 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. Les volets constitution, enrichissement et diffusion de corpus constituent donc, là aussi, une base de travail unique et de grande valeur pour la communauté du domaine.

2.3. Intérêt du point de vue culturel et pédagogique

La diffusion de données sur le 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 nos langues, l'existence de ressources fiables et finement décrites est fondamentale. Si cette question est aujourd'hui bien traitée dans le cas de documents édités par le biais du dépôt légal, 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.).

Du point de vue de l'enseignement et de l'apprentissage des langues, l'existence de données bien décrites incluant 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 d'enseignement à une époque où la référence à des « documents authentiques » a enfin supplanté les « exemples construits » ou « exemples d'école » (Duda et 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.

2.4. 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

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

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 et qui ont besoin de données de qualité et de grande taille pour développer leurs produits. La plupart des entreprises du domaine, start-up et PME, ne peuvent se permettre, vu les coûts d'investissement à prévoir, d'élaborer des ressources linguistiques à large couverture. Une telle infrastructure devrait permettre aux industriels de tester des ressources, lors des phases 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-up ou PME.

3. 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.

3.1. 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. 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[7], LMF[8] et SYNAF[9].

Un second objectif a été 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 de textes et utilisable sur tout corpus de langue écrite, l'enrichissement d'un lexique

[7] www.tei-c.org/
[8] http://www.lexicalmarkupframework.org/
[9] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37329

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

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 développement d'une grammaire couvrante du français et enfin la normalisation de divers corpus parmi lesquels COLAJE, l'Est Républicain, ESLO, PFC, TCOF[10].

3.2. Pérennisation des ressources

Afin d'assurer la pérennisation des ressources, nous avons mis en œuvre quatre 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 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.

3.3. 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, ORTOLANG collecte aussi des statistiques et assure la diffusion de notifications d'usage auprès des déposants.

4. 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 faire nous avons développé un flux de travail qui se décompose en 5 étapes.

[10] https://www.ortolang.fr/market/corpora

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

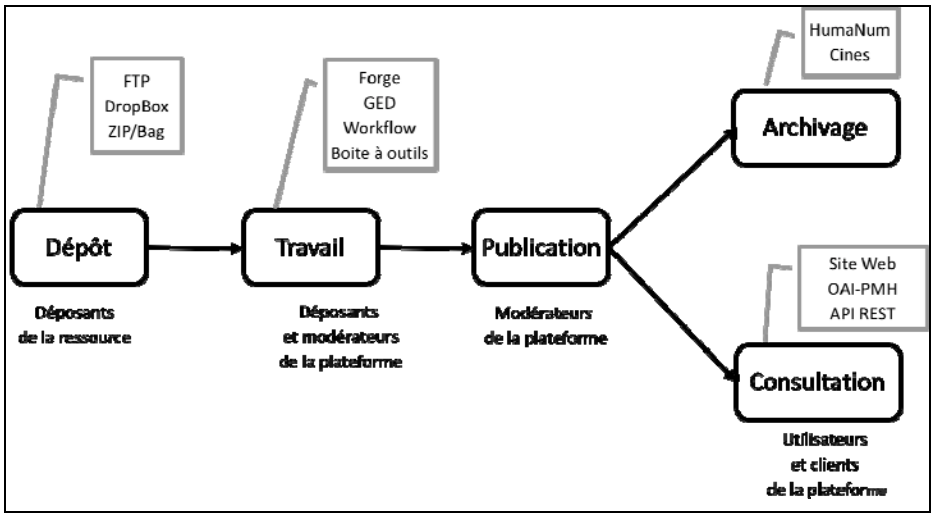

Figure 1 : Flux de travail de dépôt d'une ressource dans ORTOLANG.

4.1. 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.

4.2. 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. 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 (LPL/SLDR), et le troisième vers des ressources multimodales (MoDyCo). Le producteur dispose de plus au sein de son espace de travail de divers outils en ligne lui permettant (i) de déclarer les membres de cet espace de travail qui doivent préalablement s'ouvrir un compte sur la plateforme (ii) de déposer de nouveaux contenus et spécifier pour chaque dossier ou fichier déposé leur visibilité externe.

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

Nous avons choisi de limiter à quatre les choix d'accessibilité et de visibilité des ressources : à tous ; aux utilisateurs s'étant préalablement déclarés sur la plateforme ; aux seuls membres de l'Enseignement Supérieur et de la Recherche ; ou enfin restreint aux seuls membres de l'espace de travail (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) (iii) enrichir son travail (conversion de format, alignement, annotation, etc.) et en particulier définir les métadonnées correspondantes (iv) enfin, une fois les métadonnées définies, obtenir une visualisation de sa ressource.

4.3. 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.

4.4. 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[11] sur les parlers 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 effort financier et humain supérieur à celui de son archivage pérenne.

4.5. 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.

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

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

5. 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 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, 1 501 utilisateurs s'étaient créé un compte sur la plateforme. De plus, au cours des derniers mois nous avons été contactés par plusieurs laboratoires et divers projets qui souhaitent héberger au sein d'ORTOLANG l'ensemble de leurs ressources et outils. Nous pensons que ces données chiffrées sont autant d'indices de réussite de notre projet 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 qui ont permis la réalisation de ce projet.*

BIBLIOGRAPHIE

Barlow, Michael, Suzanne Kemmer, 2000 *Usage Based models of language*, University of Chicago Press.

Duda, Richard, Henry Tyne, 2012, « Authenticity and Autonomy in Language Learning », *Bulletin Suisse de Linguistique Appliquée*, 92, p. 86–106.

Habert, Benoit, Adeline Nazarenko, André Salem, 1997 *Les linguistiques de corpus*. Paris : Armand Colin.

Laks, Bernard, 2008, « Pour une phonologie de corpus », *Journal of French Language Studies*, 18, n° 1, p. 3–32.

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)

Pierrel, Jean-Marie, 2014, « 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, p. 169–190.

Pierrel, Jean-Marie, Etienne Petitjean, 2007, « 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, p. 327–330.

Wittenburg, Peter, Nuria Bel, Lars Borin, Gerhard Budin, Nicoletta Calzolari, Eva Hajicova, Kimmo Koskenniemi, Lothar Lemnitzer, Bente Maegaard, Maciej Piasecki, Jean-Marie Pierrel, Stelios Piperidis, Inguna Skadina, Dan Tufis, Remco Van Veenendaal, Tamas Váradi, Martin Wynne, 2010, « Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure », *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valetta, Malte, www.lrec-conf.org/proceedings/lrec2010/.

AN INFRASTRUCTURE OF LONG-TERM PRESERVATION, MUTUALIZATION, AND VALORISATION OF LANGUAGE RESOURCES: THE ORTOLANG EQUIPEX

Abstract

ORTOLANG (Open Resources and Tools for Language: www.ortolang.fr) is a French infrastructure implemented in the framework of the "Programme d'Investissement d'Avenir" (PIA) funded by the "Investissements d'Avenir" French Government program. Based on the existing resource centers CNRTL (www.cnrtl.fr) and SLDR (http://sldr.org/), this infrastructure aims to ensure the management, mutualization, dissemination and long-term preservation of language resources such corpus, dictionaries, lexicons and language processing tools, with particular focus on the languages of France. It will be used as a technical language platform of written and oral language forms, as a support of actions of coordination carried out by the TGIR Huma-Num (http://www.huma-num.fr/).

BDD-A30848 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 18.97.9.169 (2026-09-10 13:51:45 UTC)