Ortolang : Une infrastructure de mutualisation de ressources linguistiques écrites et orales

Jean-Marie Pierrel

Ortolang

Une infrastructure de mutualisation de ressources linguistiques écrites et orales

Jean-Marie Pierrel
ATILF UMR 7118, Université de Lorraine & CNRS

Résumé

Dans cet article nous présentons l'infrastructure Equipex Ortolang[49] *(Open Resources and Tools for LANGuage / Outils et Ressources pour un Traitement Optimisé de la LANGue : www.ortolang.fr) en cours de mise en place dans le cadre du Programme d'Investissements d'Avenir (PIA) lancé par le gouvernement français.*

S'appuyant entre autres sur l'existant des centres de ressources CNRTL (Centre National de Ressources Textuelles et Lexicales : www.cnrtl.fr) et SLDR (Speech and Language Data Repository : http://sldr.org/), cette infrastructure a pour objectif d'assurer la gestion, la mutualisation, la diffusion et la pérennisation de ressources linguistiques de type corpus, dictionnaires, lexiques et outils de traitement de la langue, avec une focalisation particulière sur le français et les langues de France.

*Après avoir rappelé les motivations d'un tel projet, son originalité et son caractère novateur, nous présenterons les principales caractéristiques d'*Ortolang*, ses objectifs et ses missions, l'infrastructure logicielle et matérielle de la plateforme puis les moyens mis en œuvre, avant de conclure en indiquant comment suivre et contribuer au projet.*

Mots-Clés

Ortolang, plateforme, mutualisation, corpus, ressources linguistiques

[49] *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

Abstract

This paper presents the infrastructure for the Equipex Ortolang *(Open Resources and Tools for LANGuage / Outils et Ressources pour un Traitement Optimisé de la LANGue : www.ortolang.fr) which is currently being developed as part of the French government's Investments for the Future programme.*

Drawing on existing resources such as the CNRTL (Centre National de Ressources Textuelles et Lexicales: www.cnrtl.fr) and SLDR (Speech and Language Data Repository: http://sldr.org/), the infrastructure is designed for the long-term management, sharing and dissemination of linguistic resources including corpora, dictionaries, lexicons and language processing tools, with a particular focus on French and other languages in France.

The paper briefly presents the rationale behind such an original and ground-breaking project, then describes the main characteristics and goals of Ortolang*, the platform hardware and software as well as the means available, before concluding with planned future developments and an invitation to contribute to the project.*

Keywords

Ortolang, platform, sharing, corpus, linguistic resources

1. Pourquoi une telle infrastructure ?

Dans notre société de l'information, seules les langues fortement outillées et modélisées, permettant des traitements automatiques, ont 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, les jeux sont loin d'être faits (Union Latine, 2008). Il paraît donc important et urgent de doter le français des outils indispensables à son traitement automatique, si nous souhaitons qu'à l'avenir il continue 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.

Une rapide analyse de l'évolution des sciences du langage et du traitement automatique des langues (TAL) au cours des trente dernières années montre 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 al*., 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.

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

- aux travaux de recherche pour lesquels la notion de corpus d'étude et de ressources est incontournable 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 de France s'impose donc aujourd'hui 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 mise au point d'outils d'analyse (morphologique, morphosyntaxique, lexicale, syntaxique et sémantique) est important. C'est 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 regarder le nombre de requêtes servies aujourd'hui par un site tel que le CNRTL (http://www.cnrtl.fr/aide/stat/) : plus de 450 000 requêtes par jour, venant du monde entier, sur le lexique du français !

2. Originalité et caractère novateur du projet d'équipement

Ortolang (*Open Resources and TOols for LANGuage* / Outils et Ressources pour un Traitement Optimisé de la LANGue : www.ortolang.fr), projet validé dans le cadre du programme d'investissements d'avenir (PIA), se donne comme 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. Une telle infrastructure doit permettre à la communauté de franchir un pas décisif, aujourd'hui encore à peine ébauché. Il s'agit, non seulement du contenu et de la variété des données ou outils disponibles (qui seront encore enrichis et améliorés pendant le déroulement du projet), mais aussi et surtout d'assurer la diffusion de standards clairs, internationalement reconnus, afin de pouvoir les rendre accessibles et permettre le partage, la réutilisation et la complémentation des informations. 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 langage 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 & 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, telles 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. 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 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. 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 standard 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 un 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 *et al.*, 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 aux informations radiophoniques (ESTER[50]) et aux émissions de télévision (ETAPE[51]), l'ambition unanimement affichée consiste à diversifier les styles de parole et à ouvrir les évaluations aux enregistrements de réunions (*Meetings*) et aux conversations spontanées (*Switchboard*), comme cela a déjà été réalisé aux Etats-Unis par le NIST[52] dans le cadre des campagnes RT[53]. Le projet *Ortolang* permettra 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[54] 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[55].

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

2.3. 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'Emile 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

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

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

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

[53] RT : Rich Transcription Evaluation Project, http://www.itl.nist.gov/iad/mig//tests/rt/.

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

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

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/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 en apprentissage du français langue seconde ou 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 connaissance, 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 qui servent de support au 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 devrait 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.

3. Principales caractéristiques d'*Ortolang*

3.1. Une ouverture pluridisciplinaire forte

Pour structurer le projet *Ortolang*, nous avons choisi de créer un consortium constitué de laboratoires et de centres de ressources en charge de la définition du projet et possédant des compétences complémentaires dans les domaines suivants :

- les sciences du langage à travers l'ATILF[56], le LPL[57], MoDyCo[58] et le LLL[59] ;
- l'informatique avec le LORIA[60] et l'INIST[61], mais aussi en partie l'ATILF et le LPL, laboratoires SHS d'interface avec l'informatique ;
- la maîtrise de bases de données et de l'accès à de l'information scientifique, à travers l'INIST, ainsi qu'à des ressources linguistiques, au travers de deux centres de ressources, le CNRTL[62] et le SLDR[63].

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

[56] Analyse et Traitement Informatique de la Langue Française, UMR Université de Lorraine - CNRS, www.atilf.fr.

[57] Laboratoire Parole et Langage, UMR Aix Marseille Université – CNRS, www.lpl-aix.fr.

[58] Modèles, Dynamiques, Corpus, UMR Université Paris Ouest Nanterre La Défense – Université Paris Descartes – CNRS, www.modyco.fr.

[59] Laboratoire Ligérien de Linguistique, Université d'Orléans – CNRS, www.lll.cnrs.fr.

[60] Laboratoire Lorrain de Recherche en Informatique et Applications, UMR Université de lorraine – CNRS – INRIA, www.loria.fr.

[61] Institut de l'Information Scientifique et Technique, UPS CNRS, www.inist.fr.

[62] Centre National de Ressources Textuelles et Lexicales, www.cnrtl.fr.

[63] Speech and Language Data Repository, http://sldr.org.

linguistique (MoDyCo, LPL et ATILF), linguistique expérimentale et/ou appliquée (LPL, ATILF), production et perception du langage (LPL, MoDyCo), études diachroniques (ATILF, LLL), sociolinguistiques (LLL, MoDyCo), traitement automatique des langues (LORIA, LPL, ATILF), écrit (ATILF, MoDyCo), oral (LPL, LLL, MoDyCo).

Cette proposition s'appuie sur une importante expérience acquise des équipes proposant cet équipement d'excellence. À titre illustratif, nous explicitons ci-dessous quelques atouts tant en termes de ressources et outils déjà proposés que d'insertion nationale et internationale :

- l'acquis des partenaires, centres de ressources (CNRTL et SLDR) et laboratoires qui alimenteront la version initiale de la plateforme avec un ensemble de ressources et d'outils déjà disponibles en leur sein et dont les compétences recouvrent les trois principaux aspects visés : l'oral, l'écrit et la patrimonialisation des parlers de France ;
- l'implication et la cohérence avec la Très Grande Infrastructure de Recherche (TGIR) HumaNum[64] : nous sommes partie prenante des consortiums CORPUS Ecrit[65] et IRCOM[66] ;
- l'implication et la cohérence avec l'infrastructure européenne CLARIN[67] au sein de laquelle nous avons travaillé dès la phase préliminaire ;
- la cohérence avec les efforts menés par la DGLFLF[68] et la BNF[69] sur les aspects patrimonialisation des parlers de France.

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

La plateforme *Ortolang* ne se veut être qu'une infrastructure de mutualisation pour la gestion, la pérennisation et la diffusion de corpus et d'outils sur la langue, ces derniers restant bien

[64] http://www.huma-num.fr/.

[65] http://corpusecrits.huma-num.fr/.

[66] http://ircom.corpus-ir.fr.

[67] www.clarin.eu.

[68] Délégation Générale à la Langue Française et aux Langues de France, www.dglf.culture.gouv.fr.

[69] Bibliothèque Nationale de France, www.bnf.fr.

entendu propriété des déposants (chercheurs ou laboratoires). Nous avons, de plus, prévu des moyens pour aider des laboratoires à finaliser et à normaliser leurs ressources.

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

- le respect de la charte éthique *Big Data*[70], 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 de valorisation contractuelle ;
- moyennant royalties auprès des propriétaires des ressources, dès qu'il y a valorisation contractuelle.

C'est dans cet esprit que divers contacts avec des partenaires externes ayant déposé ou souhaitant déposer leurs ressources sur *Ortolang* ont déjà été mis en œuvre. Parmi ces partenaires, on peut déjà noter :

- les consortiums Ecrit et IRCOM d'HumaNum en linguistique au travers d'appels à projets communs pour la finalisation et la standardisation de corpus ;
- les fédérations de recherche en linguistique ILF (Institut de la Langue Française) et TUL (Typologie et Universaux Linguistiques). Ainsi *Ortolang* sert de support à l'initiative "Corpus de référence du français"[71] lancée par l'ILF ;
- le laboratoire d'Informatique de Tours pour un projet d'annotation en entités nommées des corpus gérés par *Ortolang*;
- divers projets ANR (ORFEO, TermITH, OTIM, etc.) ;
- et les labex EFL[72] (Paris), BLRI[73] (Aix-Marseille).

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

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

[72] www.labex-efl.org/.

[73] http://www.blri.fr/.

4. Objectifs et missions de cette infrastructure

Les objectifs et missions du projet *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 données

L'une des difficultés actuelles pour repérer et accéder à des ressources (corpus, dictionnaires, lexiques et outils de traitement) sur notre langue réside tout à la fois dans leur grande dispersion (il n'est pas aisé de savoir quelles ressources sont disponibles et à quels endroits elles sont accessibles) et leur forte disparité, en particulier en termes de codage. 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 :

- *le catalogage des ressources et outils existants* à travers un ensemble de métadonnées normalisées. Cette action est menée en étroite coopération avec les consortiums Ecrit et IRCOM de la TGIR HumaNum ;
- *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 telles XML[74], TEI[75], LMF[76] et SYNAF[77] (Declerck, 2006) ;
- *l'enrichissement de ressources et d'outils*. Cette action s'appuie sur les équipes porteuses d' *Ortolang* (ATILF, LPL, LORIA, MoDyCo et LLL) 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[78], l'amélioration de la couverture temporelle d'un lemmatiseur du français[79]

[74] http://xml.chez.com/.

[75] http://www.tei-c.org.

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

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

[78] www.cnrtl.fr/lexiques/morphalou/.

[79] http://www.atilf.fr/dmf/LGeRM/.

et sa mise à disposition sous forme de Web Service, le développement d'outils de segmentation de phrases multilingues, le développement d'un outil d'aide à la transcription de corpus oraux, le développement de plug-ins assurant l'interopérabilité entre les différents outils d'édition et d'annotation, le développement d'une grammaire couvrante du français, la normalisation de divers corpus parmi lesquels COLAJE[80], EMERGRAM[81], l'Est Républicain[82], ESLO[83], PFC[84], TCOF[85].

De plus, afin de créer un mouvement de mutualisation largement ouvert vers des équipes externes au consortium, nous avons prévu des financements pour des appels à projets communs avec les consortiums Ecrit et IRCOM afin de prendre en charge les nécessaires travaux de normalisation des corpus et/ou outils que les équipes externes au consortium souhaitent déposer sur la plateforme *Ortolang*. C'est en particulier ce qui a déjà démarré pour le corpus CoMeRe[86] de communication médiée par les réseaux.

4.2. Pérennisation des ressources

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

- la curation des ressources et des outils ;
- un stockage sécurisé et une maintenance des ressources ;
- un archivage pérenne, à travers la solution mise en place par la TGIR HumaNum en lien avec le CINES.

4.3. Diffusion

Enfin, pour assurer la nécessaire diffusion et exploitation de ces ressources, nous prévoyons une aide et un accompagnement des utilisateurs pour la mise en place des procédures permettant à des utilisateurs de la plateforme d'exploiter ces ressources et outils mutualisés en

[80] http://colaje.scicog.fr/.

[81] http://emergram.scicog.fr/corpora.php.

[82] http://www.cnrtl.fr/corpus/estrepublicain/.

[83] http://www.lll.cnrs.fr/eslo-1.

[84] http://www.projet-pfc.net/index.php.

[85] http://www.cnrtl.fr/corpus/tcof/.

[86] http://corpuscomere.wordpress.com/.

nous appuyant sur l'expérience des centres de ressources CNRTL (www.cnrtl.fr) et SLDR (http://sldr.org/) appelés à terme à se fondre au sein d' *Ortolang*.

5. Architecture logicielle

L'architecture logicielle d'*Ortolang* s'appuie sur un centre de diffusion que l'on souhaite pleinement compatible avec les recommandations du projet d'infrastructure européenne CLARIN pour ces centres de ressources (Wittenburg *et al*., 2010) et de "centres thématiques" directement accessibles par les utilisateurs afin de permettre la navigation dans les collections de ressources ou l'obtention de ressources via des requêtes sur les métadonnées.

5.1. Un centre de diffusion compatible CLARIN

Couche basse de l'architecture logicielle d'*Ortolang*, ce centre de diffusion devra supporter des contraintes de qualité de service (disponibilité maximale) et de gestion des documents permettant d'obtenir le DSA (*Data Seal of Approval*[87]). Ce centre, entrepôt OAI-PMH, peu visible des utilisateurs, sera un dépôt fiable des données. Les fonctionnalités prévues à ce niveau sont :

- l'identification de chaque ressource par un identifiant pérenne (ou Handle) ;
- une preuve d'intégrité de la donnée associée à un identifiant pérenne devra être fournie sous forme d'une somme de contrôles liée à l'identifiant pérenne ;
- la gestion de versions. Toute modification d'une donnée doit conduire à une nouvelle version (non nécessairement préservée à long terme). Cette gestion des versions s'effectue à travers une relation dédiée dans les métadonnées ;
- l'authentification des utilisateurs à travers un mécanisme de signature unique (*Single Sign On*) lors de la consultation de données à accès restreint ;

[87] http://datasealofapproval.org.

- l'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.

De plus, concernant l'interopérabilité des métadonnées, nous serons amenés à compléter le format CMDI[88] proposé par CLARIN sur deux points :

- définition d'un schéma minimal commun faisant intervenir la notion de déposant et la notion de contrôle d'intégrité ;
- définition d'un mécanisme d'aplatissement de la structure de métadonnées permettant d'obtenir différentes vues thématiques de la même métadonnée.

5.2. Des centres "thématiques"

Partie émergée de l'équipement directement visible pour les utilisateurs, trois centres thématiques seront proposés, orientés respectivement vers les aspects textuels, oraux et patrimoniaux.

L'enjeu est de rendre visible l'ensemble des données hébergées dans le centre de diffusion à partir de chaque centre thématique. Pour autant, des centres d'intérêt différents amènent à proposer des méthodes de navigation dans les métadonnées, un filtrage d'une partie des métadonnées et des interfaces de recherche et de visualisation qui leur sont spécifiques.

Les centres thématiques n'ont pas pour vocation d'héberger des ressources, autres que celles en cours de définition. Ils moissonnent le centre de diffusion afin de disposer des métadonnées de l'ensemble des dernières versions des ressources. L'accès aux données elles-mêmes se fait alors par un renvoi vers ce centre. Les centres thématiques sont également les interlocuteurs des déposants. Il est de leur responsabilité de mettre en forme données et métadonnées avant transmission au centre de diffusion. Le centre de diffusion, quant à lui, est responsable de la mise en forme vis-à-vis des opérateurs d'archivage. Les centres thématiques doivent aussi permettre aux chercheurs de se constituer des corpus de travail de façon transparente. Ils offriront trois modes d'identification des ressources : une navigation par collection, une interface simple de recherche dans les métadonnées et une interface complexe de recherche à facette.

[88] http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml.

6. L'infrastructure matérielle mise en place

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 *Ortolang* (serveurs, système d'exploitation, disques durs, robotique de sauvegarde) et des moyens INIST partagés (réseau, pare-feu, hyperviseur, réseau de stockage et de sauvegarde (SAN), serveur de pilotage des sauvegardes, salles machines).

Dans la phase de construction du projet (2013-2016), ces moyens informatiques sont aussi utilisés pour héberger les environnements de développement.

Une double logique a été utilisée pour concevoir l'architecture *Ortolang* :

- une mutualisation des moyens entre les différents centres (diffusion/thématique ; production/pré-exploitation/développement) avec utilisation de moyens complémentaires apportés par l'INIST ;
- une architecture hautement sécurisée et à haute disponibilité.

L'infrastructure *Ortolang* est répartie dans 2 salles informatiques climatisées, sécurisées (sécurité électrique et sécurité d'accès) et avec des connexions redondantes (double attachement réseau, double attachement vers le stockage et les moyens de sauvegardes).

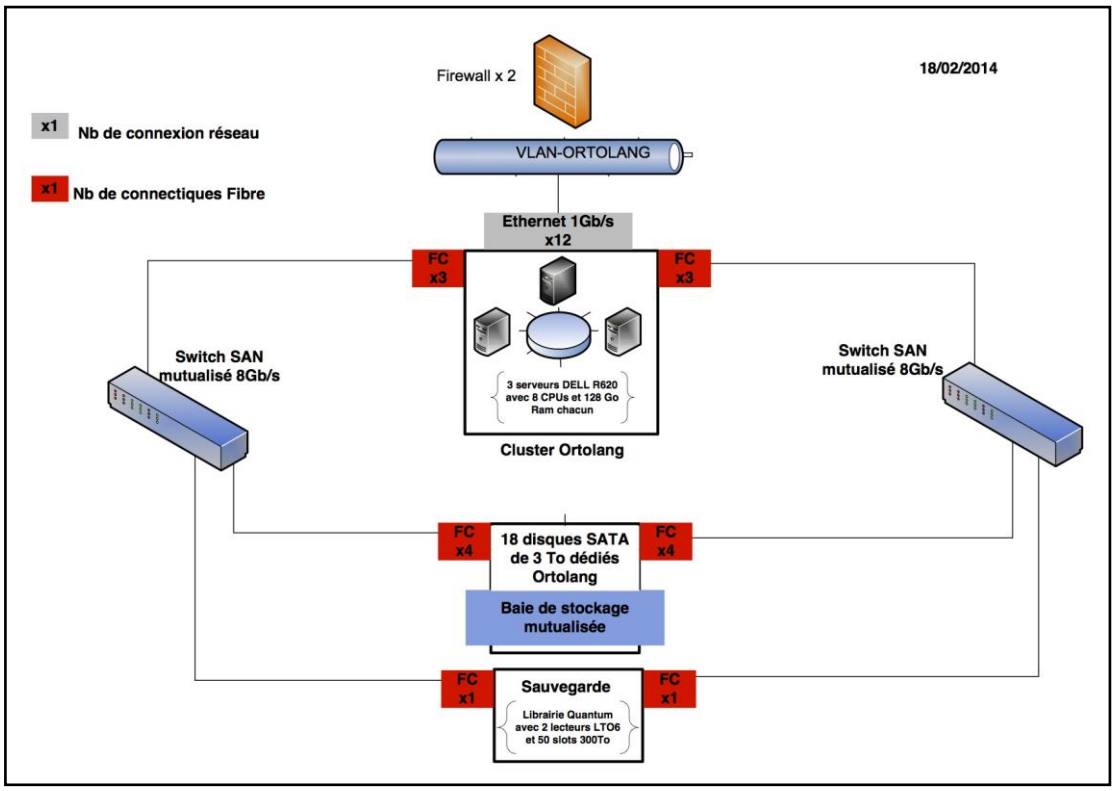
Coté logiciel, il a été choisi d'utiliser des technologies de virtualisation pour avoir le maximum de souplesse et exploiter au maximum les ressources physiques (puissance CPU, capacité mémoire centrale RAM, pool de stockage). Ainsi trois serveurs physiques ont été acquis pouvant héberger de l'ordre d'une dizaine de machines virtuelles qui peuvent être déplacées d'un serveur physique à l'autre, soit pour assurer des maintenances programmées, soit pour assurer la continuité en cas de défaillance matérielle.

Les serveurs physiques sont connectés au Réseau de Stockage SAN 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. La partie stockage est configurée en utilisant des mécanismes de redondance et correction d'erreur Raid 6.

Tous les équipements ont été acquis avec une couverture de maintenance et de support jusqu'à 2016, soit la fin de la Tranche 1 du projet et une jouvence de ces matériels est prévue en 2016 et 2019. Ils regroupent :

- *des serveurs* : nous avons choisi une structure de serveurs permettant de définir des machines virtuelles et s'appuyant sur trois serveurs DELL R620 bi-processeur, disposant chacun de 128 Go de mémoire vive, sur lesquels est implanté un système de virtualisation VMware VSphere Entreprise et le système SUSE Linux Enterprise ;
- *un système de stockage* : acquisition par *Ortolang* de disques d'une capacité totale brute de 50 To insérés dans le sous-système de stockage INIST ;
- *un sous-système de sauvegarde* : acquisition par *Ortolang* d'une robotique dédiée (deux lecteurs LTO6 et une librairie pouvant contenir cinquante cartouches) et du logiciel de sauvegarde/restauration HP Data Protector. Ces équipements s'intègrent dans l'infrastructure de stockage et de sauvegarde de l'INIST (SAN) et bénéficient du serveur pilotant les sauvegardes (DELL R910) et de la licence *site* HP Data Protecteur relative à la capacité sauvegardée et aux clients de sauvegarde. La capacité de chaque cartouche étant de 2,5 To (6,2 To pour un facteur de compression de 2.5 apportée par les lecteurs LTO6), la capacité totale est donc de 125 To (312,5 To compressés).

Figure 1 – Architecture matérielle mise en œuvre pour *Ortolang*



7. Les moyens mis en œuvre pour assurer une telle mission

Dans le cadre du programme d'investissements d'avenir lancé par le gouvernement français, le projet *Ortolang* est doté tout à la fois d'une structure de pilotage transparente et d'un budget lui permettant de remplir au mieux les objectifs qu'il s'est fixés.

7.1. Structure de pilotage mise en place

Outre un comité technique, composé des représentants des divers partenaires du projet, qui assure le suivi opérationnel des actions d'*Ortolang* et qui se réunit toutes les six semaines, nous nous sommes dotés d'un comité d'orientation et d'un comité scientifique.

7.1.1. Le comité d'orientation

Ce comité d'orientation regroupe des membres ès-qualité représentant les diverses institutions partenaires d'*Ortolang* et se réunira au mois une fois tous les deux ans. Il a pour objectif d'évaluer l'activité du projet en fonction des objectifs définis et de proposer les grandes orientations pour la suite du programme.

7.1.2. Le comité scientifique

Le comité scientifique réunit à parts égales des collègues issus de la communauté scientifique nationale et de la communauté internationale. Placé sous la présidence de Laurent Romay, INRIA & DARIAH, il est composé de Dan Broeder, MPI Nijmegen et CLARIN ; Marin Dacos, Directeur du Centre for Open Electronic Publishing ; Cédric Fairon, UCL Louvain ; Alexander Geyken, BBAW Berlin ; Benoît Sagot, INRIA, membre du LabEx EFL ; Véronique Traverso, CNRS, ICAR & LabEx ASLAN ; Andreas Witt, IDS Manheim ; le président de l'ATALA, Association pour le Traitement Automatique des Langues[89] ; le president de l'AFCP, Association Francophone de Communication Parlée[90].

7.2. Moyens financiers mobilisés

Outre l'investissement des partenaires, les moyens financiers fournis par l'ANR pour la réalisation du projet s'élèvent à un total de 2 600 k€ HT :

- 2 200 k€ pour la période de mise en place du projet (2013-2016), dont 1 530 k€ de frais de personnel pour l'élaboration de l'infrastructure logicielle, des ressources et des outils ;
- 400 k€ pour assurer le fonctionnement pour la période 2016-2019.

La pérennisation au-delà de 2020 de l'infrastructure mise en place sera quant à elle directement liée à la réussite du projet.

8. En guise de conclusion : suivre l'évolution du projet et y contribuer

La mise en place en place de la plateforme *Ortolang*, prévue sur plusieurs années, verra l'ouverture des différents services s'échelonner jusque mi-2016 au travers de versions successives de la plateforme suivant le planning de réalisation présenté Figure 2, planning dont on trouvera une mise à jour régulière sur le site du projet.

La première version, opérationnelle depuis avril 2014 et accessible via le site www.ortolang.fr, permet de rechercher une ou des ressource(s) au travers d'une recherche par

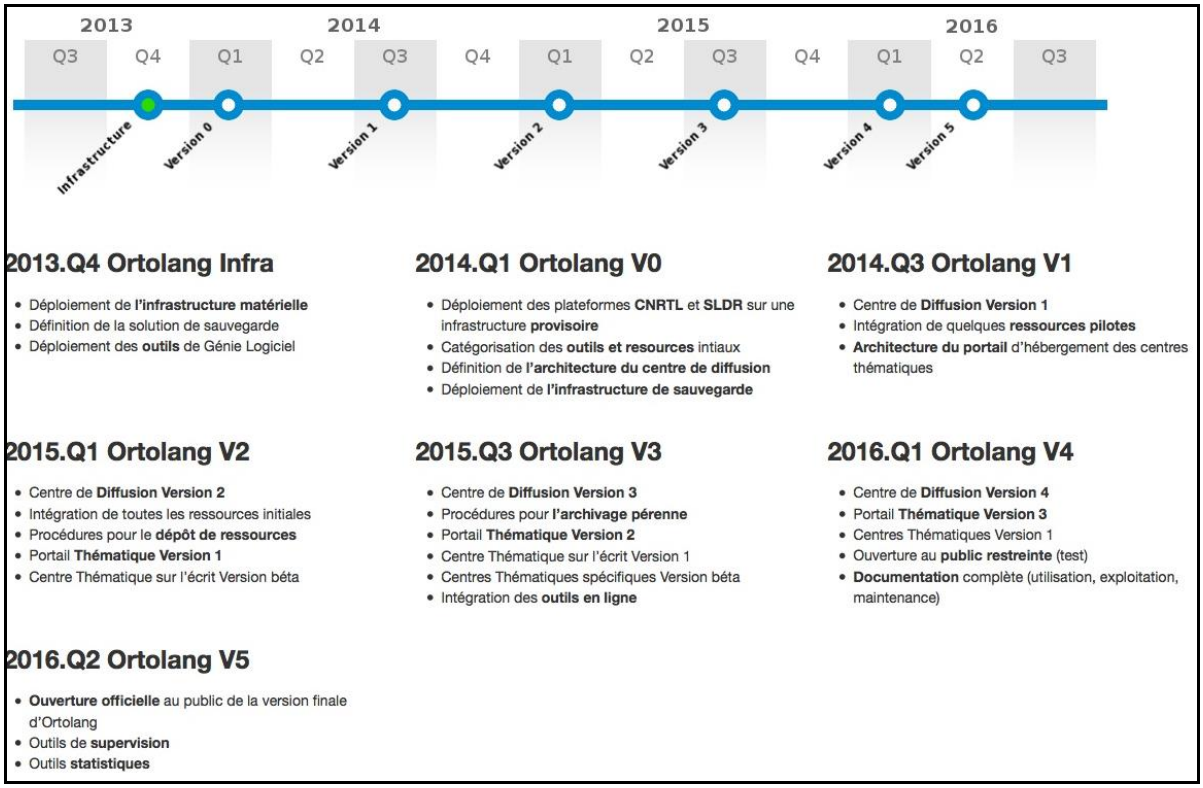
[89] http://www.atala.org/.

[90] http://www.afcp-parole.org/.

facettes permettant des sélections suivant les droits d'utilisation (libres ou sous droits), les langues, les types (corpus écrits, corpus oraux ou outils), les dates de création, les formats ou les éditeurs des ressources. Cette première version fournit aussi un accès au CNRTL et au SLLR qui seront petit à petit complètement intégrés à *Ortolang*. La version suivante accueillera début 2015 les données des centres dans le serveur de diffusion *Ortolang* afin de permettre le téléchargement des ressources directement à partir du portail. Puis apparaitra l'émergence des portails thématiques *écrit*, *oral* et *patrimonial* qui permettront de rechercher plus efficacement dans un type particulier de ressources.

Ce même site www.ortolang.fr permet de suivre l'évolution du projet et d'accéder à un ensemble de documentation sur les ressources gérées. Dès aujourd'hui, il vous est proposé de contribuer au projet au travers de dépôts de ressources en prenant contact avec l'équipe technique du projet à l'adresse : contact@ortolang.fr. Au final, la réussite d'un tel projet reposera 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.

Figure 2 – Planning de réalisation d'*Ortolang*



Références

Barlow, M. & Kemmer, S. (2000). *Usage based models of language*. Chicago : University of Chicago Press.

Declerck, T. (2006). "SYNAF : towards a standard for syntactic annotation". *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006). Genève, 22-28 mai 2006. Disponible en ligne. http://www.lrec-conf.org/proceedings/lrec2006/

Duda, R. & Tyne, H. (2012). "Authenticity and autonomy in language learning". *Bulletin Suisse de Linguistique Appliquée*, vol. 92. pp. 86-106.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin.

Haton, J.-P., Cerisara, C, Fohr, D., Laprie, Y. & Smaïli, K. (2006). *Reconnaissance automatique de la parole : du signal à son interprétation*. Paris : Dunod.

Hunston, S. & Francis, G. (2000). *Pattern grammar : a corpus-driven approach to the lexical grammar of English*. Amsterdam : John Benjamins.

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

Newmeyer, F. J. (2003). "Grammar is grammar and usage is usage". *Language*, vol. 79. pp. 682-707.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

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

Union Latine. (2008). *Langue et cultures sur la toile : enquête 2007*. Disponible en ligne. http://dtil.unilat.org/LI/2007/index_fr.htm

Wittenburg, P., Bel, N., Borin, L., Budin, G., Calzolari, N., Hajicova, E., Koskenniemi, K., Lemnitzer, L., Maegaard, B., Piasecki, M., Pierrel, J.-M., Piperidis, S., Skadina, I., Tufis, D., Veenendaal, R. van, Váradi, T. & Wynne, M. (2010). "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. Disponible en ligne. http://www.lrec-conf.org/proceedings/lrec2010/index.html

À propos de l'auteur

Jean-Marie Pierrel est Professeur à l'Université de Lorraine et membre du laboratoire Atilf. Informaticien-linguiste, il est responsable du CNRTL et directeur de l'Equipex *Ortolang*. Spécialiste de traitement automatique des langues, ses recherches portent aujourd'hui sur la définition, l'enrichissement, la gestion informatique et l'exploitation de ressources linguistiques (corpus et dictionnaires informatisés) pour l'ingénierie des langues, le traitement automatique des langues et la recherche en linguistique informatique.

Courriel : Jean-Marie.Pierrel@atilf.fr

Toile : www.atilf.fr, www.cnrtl.fr et www.ortolang.fr

Adresse : Atilf, Université de Lorraine & CNRS, 44 avenue de la Libération, BP 30687, 54063 Nancy cedex