Bilan intermédiaire du Consortium Corpus Oraux et Multimodaux de la TGIR Huma-Num

Stéphane Robert

Huma-Num ***Consortium Corpus Oraux et Multimodaux de la TGIR***

Bilan du consortium IRCOM pour l'année 2013

Sommaire :

1 - RAPPEL DES MISSIONS DE L'IRCOM ET BILAN À MI-PARCOURS

Les missions du consortium IRCOM définies par l'accord de labellisation signé le 2 mars 2012 sont les suivantes :

- *organiser et accompagner le développement de corpus oraux et multimodaux en linguistique en aidant les chercheurs à s'approprier les outils nécessaires et à développer des standards communs de référence*.

L'appropriation d'outils standards de constitution, d'annotation, d'analyse, de valorisation et d'archivage des corpus a été assurée depuis 2012 par un plan de formation qui, à l'issue de la période de labellisation, balaiera l'ensemble de la chaîne. Deux de ces quatre étapes (constitution vidéo/audio et annotation avec ELAN) ont donné lieu à 3 formations pendant la période 2012-2013. Toutes les formations sont redoublées d'une année sur l'autre de sorte que soixante personnes seront formées à chaque étape,. Au demeurant un groupe de travail est en contact étroit avec les principaux acteurs européens et nationaux qui développent, promeuvent et utilisent des standards de métadonnées, d'annotations et d'archivage. L'interopérabilité des outils et la réutilisabilité des corpus sont les deux axes de travail de ce groupe qui œuvre dans ce sens en relation avec le consortium Corpus écrits et avec les acteurs nationaux du domaine. En 2012, une session de formation en commun avec deux autres consortiums — Corpus écrits et CAHIER — sur le balisage en TEI a d'ailleurs permis à plusieurs participants de se perfectionner.

- *contribuer à développer la valorisation, la visibilité et l'accessibilité des fonds existants, aider à améliorer leur mise à disposition et leur interopérabilité afin d'intégrer les réseaux internationaux*.

L'année 2012 a permis d'établir un inventaire fin des corpus oraux et multimodaux (finalisés ou non) et de dresser la liste des besoins des producteurs de corpus (voir rubrique Corpus sur le site de l'IRCOM : inventaire et questionnaire en ligne pour documenter un corpus). Un programme d'aide technique et financière à la finalisation de corpus, lancé en 2013, a permis de contribuer à l'amélioration de l'état des corpus incluant la documentation, la conversion de formats, la normalisation et le dépôt auprès de centre de ressources existants. Ce programme d'aide devrait être reconduit en 2014. Un glossaire constitué en 2012 et mis en ligne (site IRCOM) contenant 63 entrées validées par le comité scientifique et le comité de pilotage couvre assez largement le champ de la terminologie de la linguistique de corpus ainsi que des humanités digitales associées. Il sera enrichi dans les mois qui viennent, c'est pour l'instant une base d'entrées parcellisées mais il répond à des attentes ponctuelles et il constituera une base définitoire de référence pour la constitution d'un manuel à l'usage des producteurs de corpus.

▪ *contribuer à la structuration de la communauté des producteurs et utilisateurs de corpus oraux et multimodaux dans ces pratiques et réflexions.*

Un site (http://ircom.corpus-ir.fr/site) et un wiki associé (http://www.ircom.corpus-ir.fr/wiki/) permettant à la fois de contribuer de manière collaborative et de valoriser le travail auprès de la communauté a été mis en place en 2012 et mis en ligne début 2013. Plusieurs journées d'étude en 2012 et en 2013 ont permis de diffuser l'information (réunion ouverte organisée par le comité scientifique), de sensibiliser un public (journées d'étude communes à plusieurs consortiums notamment sur les questions de droits), d'approfondir les réflexions méthodologiques et scientifiques (journées d'étude organisées par les groupes de travail 1, 2, 4, 5 et 6). Au demeurant un groupe de travail (groupe 5 Questions juridiques et éthiques), en étroite collaboration avec les consortiums Corpus écrits et CAHIER, a commandité un rapport sur des questions juridiques à deux laboratoires de droit qui devrait être remis à la fin de l'année.

Résumé des modalités d'action de l'IRCOM en 2012-2013 :

- Concertation avec la communauté et travaux des groupes de travail de l'IRCOM
- Formations et soutiens à des actions de formation
- Veille scientifique et technologique
  - Recensement des corpus existants et de leurs besoins
  - Mise en place d'un Wiki – espace de travail et de ressources
  - Transformation du Wiki en Site public
- Aide à la finalisation et pérennisation de corpus

2 - ORGANISATION ET MODE DE FONCTIONNEMENT DE L'IRCOM ET RETOUR À MI-PARCOURS

Le consortium IRCOM a choisi de se doter d'un comité scientifique qui le conseille. Les membres de ce dernier, en étroite collaboration avec ceux du comité de pilotage, animent des groupes de travail thématiques. L'activité des groupes est fonction de leurs thématiques, certains font œuvre de formation pour la communauté, d'autres affinent techniquement une standardisation et travaillent sur l'interopérabilité, d'autres encore ouvrent un chantier juridique et éthique essentiel. Les temps de valorisation et de retour varient d'un groupe à

l'autre. Les thématiques se complètent et parfois se succèdent, aussi l'avancée des activités est-elle parfois alternée. Les activités de l'un des groupes de travail (groupe 3 Corpus multilingues et corpus plurilingues) ont dû être suspendues en 2013, faute de répondants au sein du Conseil scientifique, un nouvel appel à participation sera lancé lors d'une prochaine journée de concertation prévue en février 2014.
Les besoins exprimés par la communauté (enquête 2012) montrent que l'information quant à l'hébergement et l'archivage pérenne des corpus est diffuse et parfois opaque pour les utilisateurs. Cette thématique finaliste pour les producteurs de corpus mais cruciale pour leur standardisation a donné lieu, fin 2013, à la création d'un sixième groupe de travail sur l'hébergement et l'archivage pérenne des corpus oraux (groupe 6, voir infra).
Afin de faire mieux connaître les activités du consortium auprès de la communauté et pour renouveler une partie des membres du comité scientifique, très sollicités, une réunion de mi-parcours sera organisée en février 2014.
Au seuil de la deuxième partie de la période de labellisation, le comité de pilotage exprime une certaine satisfaction pour le travail accompli par les membres de l'IRCOM mais également, au-delà d'une expérience effective de partage de travaux avec d'autres consortiums, il tient à rappeler, précisément à l'aune de ces partages, la singularité des terrains, des méthodes, des outils, des corpus en eux-mêmes et de leurs annotations qui caractérise son champ d'investigation : oral et multimodal. Ces spécificités réclament de la continuité, elles demandent toute l'attention des membres d'HUMA-NUM.

Résumé de l'organisation de l'IRCOM :

- Un Comité de pilotage (10 personnes)
- Un Conseil scientifique (30 personnes + le CP)
- Des groupes de travail en interaction avec la communauté :
    1. Corpus linguistiques et finalités scientifiques
    2. Interopérabilité
    3. Corpus multilingues et corpus plurilingues *(activités suspendues en 2013)*
    4. Multimodalité et modalité visuo-gestuelle
    5. Questions juridiques et éthiques, droits des personnes et des producteurs de corpus
    6. Hébergement et archivage pérenne des corpus oraux *(création octobre 2013)*

3 - BILANS DES GROUPES DE TRAVAIL

Groupe de travail 1 : Corpus linguistiques et finalités scientifiques

Objectifs
L'objectif de ce groupe de travail est de faire émerger les questions scientifiques posées par la multiplication des corpus oraux et multimodaux, et l'importance grandissante de leur place dans l'élaboration et la validation d'hypothèses linguistiques.
Les questions qui animent les réflexions du groupe sont les suivantes:
1- Qu'est-ce qu'un corpus de linguistique ? (en incluant: Qu'entend-on par données primaires? Données secondaires ? et peut-on définir ce qu'est un "corpus minimal"?)

2- Y-a-t-il des possibilités de transferts méthodologiques entre les corpus appartenant à différents champs des sciences du langage ?
3- Sous quelles conditions la réutilisation de corpus déjà annotés à des fins de recherche différentes est-elle possible ?
4- Sous quelles conditions la comparaison de corpus réalisés par des producteurs différents est-elle possible (en dehors des questions techniques d'interopérabilité logicielle) ?
5- Bilan rétrospectif des corpus ayant permis des avancées importantes dans les sciences du langage (et état des lieux des manques)
6- Le rôle des traitements informatiques sur le changement des types d'analyse menées sur les corpus
7- "Un corpus de A à Z": recommandations et bonnes pratiques
8- Quelles métadonnées ? Pour quels usages ?
9- Incubateur de projets: quels questionnements innovants peuvent étendre et renouveler les perspectives sur les corpus oraux et multimodaux ?

Les activités de ce groupe de travail reposent pour une part importante sur les résultats établis par les autres groupes de travail. Aussi, il devrait prendre sa pleine mesure dans la deuxième partie de la période de labellisation.

Activités en 2013
Organisation d'une journée d'étude le 9 décembre 2013 autour *d'Une terminologie de base pour comparer les descriptions de langues et pour élaborer les outils de traitement automatique de corpus*.

▪ **Objectif de la journée**
Cette journée proposée sur une initiative de José Deulofeu appuyée par des membres du groupe de travail Ircom-GT1 a pour objectif d'avancer dans la réponse aux questions suivantes posées dans le texte de référence du Groupe de Travail N°1 de Ircom Corpus oraux :

> « Ainsi, en dehors des questions techniques d'interopérabilité, peut-on mettre en regard des corpus de différentes langues afin de mener des études typologiques ou comparatives ? Quelles annotations, et quelles métadonnées, sont nécessaires pour la comparaison ? Ces questions peuvent déboucher sur un travail d'enrichissement de la base ISOCat (terminologie linguistique) ».

▪ **Attendus**
Une réponse à ces questions est urgente. La communauté scientifique commence en effet à disposer de nombreux corpus outillés, y compris pour le français, incluant les variétés d'oral et d'écrit spontané. De sorte que la question, jusqu'ici sans réponse, de la validation des données sur lesquelles sont fondées les études est en voie de résolution : on pourra désormais, au moins en grande partie, vérifier les sources sur lesquelles sont fondées les analyses. Les avancées parallèles de la « linguistique expérimentale » complèteront cette entreprise visant à donner à notre discipline une base empirique contrôlable.
Mais l'exploitation de ces ressources ne semble pas avoir fait l'objet d'aussi nettes avancées, si l'on considère les cadres dans lesquels les linguistes présentent leurs analyses. Ces cadres sont à la fois divers et mal définis, de sorte que la mise en commun des résultats est souvent

difficile. Deux cas de figure au moins se présentent. Soit le chercheur utilise un cadre particulier (G binding, lexical functional grammar, construction grammar, HPSG...) et le lecteur doit en maitriser l'essentiel pour tirer parti des résultats de l'analyse. Soit il utilise un cadre plus ou moins démarqué de la tradition grammaticale scolaire de telle ou telle langue et les analyses reposent alors sur beaucoup d'implicite (Que met-on sous les termes prédicat ou circonstant par exemple ?), de sorte que beaucoup de malentendus sont possibles dans la lecture des résultats, ce qui est un obstacle à l'accumulation des connaissances. Pour les analyses utilisant des corpus outillés, s'ajoute la question de la relation entre les notions utilisées par les linguistes et celles qui sont utilisées par les spécialistes de TAL dans l'élaboration des systèmes d'annotation et des outils d'analyse automatique.

Pour permettre un meilleur dialogue entre ces divers chercheurs et un meilleur cumul des résultats, notamment entre chercheurs sur le français et chercheurs sur des « langues peu décrites », nous proposons, pour ce qui concerne la recherche d'une terminologie unifiée entre linguistes, de confronter un certain nombre de réponses.

- la proposition de Dixon : écrire les descriptions dans une "Fonds commun descriptif." Il a exposé ce projet dans deux ouvrages disponibles (Basic Linguistic theory2009-10).
- la réponse d'Haspelmath (2010), qui dans une démarche voisine a réfléchi à une "Framework-free grammatical theory" dans The Oxford Handbook of Linguistic Analysis -
- le cadre choisi par Claire Blanche Benveniste dans son propre ouvrage posthume de la collection "Les langues du monde" (Le français : aspects de la langue parlée , 2010, Peeters).
- la réponse de Frajzyngier (2013) qui propose une 'non-aprioristic typology', dans Non-aprioristic typology as a discovery tool, in Thornes, Tim, Erik Andvik, Gwendolyn Hyslop and Joana Jansen (eds.), Functional-Historical Approaches to Explanation: In honor of Scott DeLancey . 2013. John Benjamins: Amsterdam-Philadelphia (pp. 3–26)
- Les choix opérés par les spécialistes de TAL dans l'élaboration des logiciels de tagging en POS et de parsing (constituants ou relations de dépendance)

L'idée est de s'appuyer sur un examen critique de ces propositions pour construire un langage commun explicite permettant de présenter les généralisations descriptives dans la partie que la tradition anglo-saxonne appelle "facts" et qui précède la partie "analysis" dans les articles bien construits. Ces tentatives me semblent apporter un début de réponse aux questions posées plus haut. Nous proposons donc d'organiser une séance du consortium "corpus oraux" sur le thème : quelle terminologie pour les descriptions basées sur les corpus? La réponse qui se construira dans le temps constituerait un "basic descriptive framework" à la française, si j'ose dire, Cette tentative d'unifier les outils descriptifs pourrait utilement être menée parallèlement à une réflexion sur les possibles convergences qui se dessinent entre cadres formalisés. Mais ces deux questions ne peuvent être traitées dans une même journée. On pourrait en revanche aborder l'autre question : confronter ces avancées vers l'harmonisation des cadres descriptifs aux notions utilisées dans les outils de TAL pour annoter et interroger les corpus.

De fait, les outils de tagging en POS et les analyseurs syntaxiques s'appuient plus ou moins explicitement sur une terminologie largement inspirée de la grammaire scolaire. Cette stratégie pragmatique vise évidemment à faciliter la tâche des programmeurs et des utilisateurs et dans une mesure qu'il faudrait évaluer, à obtenir de bonnes performances aux tests comparatifs de réussite auxquels ils sont soumis. La prise en compte de corpus oraux et de langues hors tradition écrite scolaire est sans doute entrain de changer la donne, comme

en témoignent les entreprises de normalisation d'annotation entreprises à l'échelle internationale (ISOcat). Il serait utile d'envisager l'articulation de ces propositions à la pratique des linguistes et des spécialistes de TAL.
La discussion pourrait être centrée sur les points suivants :
- Comment distinguer dans les analyses grammaticales au sens strict relations syntaxiques et relations sémantiques (cerner par exemple les limites de la notion de prédicat, sans doute une des notions les plus toxiques à cet égard, pour la rendre opératoire). Le changement de perspective, par ailleurs légitime, dans la description des langues du niveau morphosyntaxique au niveau sémantico-pragmatique très négligé durant la période d'influence chomskienne, a eu un effet pervers. On a considéré que la question de l'analyse syntaxique des langues était secondaire et que, finalement , on pouvait se contenter d'un modèle intuitif « scolaire », sans tenir compte des critiques fondées dont il avait été l'objet durant la période structuraliste. On en revient souvent à des descriptions de type onomasiologique, notamment avec l'idée que les langues seraient des systèmes de codage d'une organisation pragmatico sémantique sous-jacente. Cette attitude n'est-elle pas une régression non justifiée par rapport aux études partant de la forme (sémasiologique) et fondées sur l'idée que l'objet propre de la linguistique est l'étude des signes ou constructions associant conventionnellement forme et sens ? Cette dernière tradition est d'ailleurs remise à jour par la perspective « construction grammar », avec des tentatives de modélisation (sign based construction grammar (SAG 2012).
- Comment mettre de l'ordre dans les notions qui étendent le domaine d'observation de la "phrase" au discours ? La perspective macrosyntaxique (Morel, Blanche-Benveniste Berrendonner…) se propose d'étendre la description des signes vers le domaine du discours de la façon la plus explicite. Quelle part d'héritage de cette perspective pourrait-on retenir dans la terminologie commune et comment l'articuler avec des propositions émanant de spécialistes de l'analyse du discours ?
- Comment articuler la description des structures (compétence) et celle de la réalisation plus ou moins « fidèle » de ces structures dans la langue orale spontanée (problème des disfluences, ruptures de construction, des « projections » et autres greffes, parenthèses et types de planification…) ?
- Comment intégrer l'annotation de la multimodalité ?

Interventions de: José Deulofeu, Amina Mettouchi, Philippe Blache,
Invitation de: Gil Francopoulo (ISOCat), Rodolfo Delmonte (Université Ca'Foscari Venise).
Participation: Florence Lefeuvre, Laurent Gauthier, Philippe Martin, Paola Pietrandrea, Anne Lacheret, Brigitte Gardia, Sylvain Kahane, Jeanne-Marie Debaisieux.

Activités 2011-12
Outre les échanges sur le Wiki et la liste de diffusion du groupe 1, ainsi que la participation à l'élaboration du glossaire, deux journées d'étude ont été organisées par ce groupe en 2012:

▪ **journée du 22 octobre**: plusieurs points étaient à l'ordre du jour et ont donné lieu à des présentations et des synthèses:
Typologie des utilisations de corpus dans les analyses linguistiques: une présentation des principaux projets, notamment ceux financés par l'ANR a été faite et leurs caractéristiques et apports synthétisés. La discussion a débouché sur la proposition d'engager les porteurs

d'ANR (et de projets corpus oraux et multimodaux déjà en ligne) à détailler davantage les méthodes et les outils d'analyse, ainsi que le type d'alignement choisi.
Corpus "hautement" enrichis, contenant des annotations dans tous les domaines linguistiques et complétées par des données physiologiques. Ce type de corpus pose en particulier le problème de l'alignement des différents niveaux d'annotation. La discussion a débouché sur des recommandations: Généraliser la description/documentation des données pour arriver à une interopérabilité opératoire. Définir pour chaque corpus 1) l'organisation et la structuration des données 2) la dépendance des données entre elles.

▪ **journée du 3 décembre**: plusieurs points étaient à l'ordre du jour et ont donné lieu à des présentations et des synthèses:
Transcription: une table-ronde a souligné l'aspect non-neutre de la transcription, et les problèmes posés par les choix retenus: orthographe, API, notations diverses, notamment en langues des signes. Des préconisations concernant l'explicitation des choix d'annotation ont été proposées.
Corpus de référence/évolutif, données primaires/secondaires: cette table-ronde a souligné l'importance de définir des notions telles que: Corpus de référence, Corpus Pilote, Corpus d'Etude, Corpus représentatif, Archive, Corpus ouvert/Fermé.
Approches Corpus-based/corpus-driven: cette table-ronde a permis de mieux définir les apports de chaque approche pour les problématiques liées aux recherches linguistiques sur les corpus oraux.
Campagnes d'annotation et accord inter-annotateurs: cette table-ronde a attiré l'attention sur les pratiques d'annotation très diverses selon les types de corpus: langues bien vs peu décrites, approches linguistiques vs TAL, etc. Les méthodes de calculs permettant de qualifier le corpus et sa cohérence ont été présentées et analysées.

Participants aux activités 2011-12: Philippe Boula de Mareuil, Paul Cappeau, Sylvain Kahane, Florence Lefeuvre, Amina Mettouchi, Brigitte Garcia, Anne Lacheret-Dujour, Damien Chabanal, Sophie Rosset, Laurence Devillers, Claudine Chamoreau, Paola Pietrandrea, Marie-Paule Péry-Woodley, Jeanne-Marie Debaisieux, Philippe Blache, Annelies Braffort, José Deulofeu, Laurent Gautier, Gudrun Ledegen, Paola Pietrandrea.

Perspectives 2014
Dans la continuité avec les travaux menés depuis 2011, le groupe de travail organisera sa réflexion 2014 autour d'un ensemble de questions donnant lieu à l'organisation d'une ou plusieurs table(s)-ronde(s):
- quelles sont les unités pertinentes à l'oral / dans les interactions multimodales, et comment par conséquent segmenter les corpus oraux et multimodaux ?
- quels contenus de métadonnées recommander, pour qu'un corpus soit moissonable, réutilisable, partageable dans des domaines d'analyse différents (p.ex. linguistique et anthropologie, etc).
- les outils de requête: où en sommes-nous ? Ils vont en effet de la simple utilisation d'expressions régulières à des systèmes de quasi programmation, notamment pour répondre à la question des requêtes multicouches (prosodie, morphosyntaxe, pragmatique) (voir le site du Natural Language Toolkit).

Groupe de travail 2 : Interopérabilité

Objectifs

Le groupe de travail 'Interopérabilité' a pour objectif de recenser les usages et les besoins des laboratoires en terme de formats et d'échanges de données de corpus oraux et multimodaux pour apporter des solutions aux problèmes que nous rencontrons au quotidien dans nos unités.

Une des clés de la diffusion et de l'usage massif des corpus de langage oral et multimodaux est la qualité de la réutilisabilité des corpus par les linguistes et les autres usagers scientifiques ou non scientifiques sans intervention technique autre que celle du travail linguistique proprement dit. Ainsi par exemple, un corpus conçu dans une recherche sur la pragmatique conversationnelle devrait pouvoir être utilisé directement pour travailler sur sa syntaxe ou sa phonologie, sans autre besoin de celui du passage d'un logiciel adapté à la première tâche à des logiciels adaptés aux autres tâches. Dans l'idéal, ce passage devrait permettre d'accroitre les informations disponibles sur un corpus et non pas de créer trois corpus parallèles incompatibles.
L'interopérabilité des corpus se heurte à la variété des usages linguistiques mais aussi à la variété des pratiques et des outils disponibles dans les laboratoires. Le but du groupe de travail « Interopérabilité » est de coordonner les réflexions et mutualiser les solutions pour les diffuser et les pérenniser dans la communauté.

Activités en 2013

▪ L'inventaire en ligne réalisé et maintenu par l'IRCOM a permis d'identifier les corpus oraux et multimodaux dans nos laboratoires, leurs formats mais aussi les pratiques qui les accompagnent pour décrire ces ressources et les enrichir par des transcriptions et des annotations. A partir de ce recueil, le groupe 'Interopérabilité' a pu identifier les métadonnées (descripteurs) et les logiciels utilisés ainsi que les difficultés rencontrées dans les unités pour exploiter ces ressources, les réutiliser au sein du laboratoire dans d'autres applications ou les partager avec d'autres unités de recherche.
Au-delà des laboratoires, le statut fédérateur du groupe de travail a permis d'engager les discussions auprès des instances de normalisation (Tei Council, Clarin ou Dariah, …) pour prendre en compte les spécificités de l'oral et permettre à nos corpus d'être représentés dans les principaux standards.

Pour atteindre cet objectif, le groupe de travail a choisi de diffuser ses avancées sur le site web de l'Ircom, essentiellement dans la rubrique "Ressources" avec :

- des informations sur les logiciels et sur les normes, leur stabilité et leur interopérabilité,
- des tutoriaux sur les logiciels d'alignement et de transcription,
- des exemples explicites de corpus oraux dans les principaux standards (Dublin Core, Tei, Imdi, …) donc interopérables avec un ensemble de logiciels massivement utilisés dans la communauté,
- des exemples de fonctions réalisables dans les principaux logiciels avec leurs formats d'entrée et de sortie,

- des solutions d'intégration des différentes briques logicielles ou de formats de description permettant d'arriver à une interopérabilité aisée pour tout corpus réalisé par un outil logiciel moderne et ouvert.

En terme d'organisation, le groupe de travail a choisi dans un premier temps de travailler sur l'alignement et l'annotation des transcriptions qui sont effectués avec différents logiciels suivant la nature des annotations, le type de signal (audio/vidéo) mais surtout le niveau de granularité de ces annotations. Les chercheurs sont en effet amenés à utiliser plusieurs outils pour finaliser leurs transcriptions et doivent gérer les pertes d'informations d'un logiciel à l'autre. Le travail en cours consiste à identifier les difficultés rencontrées pour pouvoir les contourner et dans l'idéal proposer une solution intégrée composée de l'ensemble des briques logicielles avec les passerelles appropriées. Un autre problème est celui de la description des formats (structuration des données paramétrables dans certains logiciels) et des personnes intervenants dans les corpus. Ces données participent des transcriptions et des métadonnées et sont codées de manière très variables selon les logiciels.

Les instances nationales et européennes ont été considérablement modifiées cette dernière année avec la création d'Huma-Num et du nouvel équipex Ortolang pour la France et la mise en œuvre de Dariah et de Clarin pour l'Europe. Le groupe de travail se doit de contribuer à ces instances et d'être engagé dans les discussions portant sur l'interopérabilité. A cet effet la prochaine réunion planifiée en décembre 2013 portera sur la mise en commun des formats TEI déjà opérationnels dans nos unités pour décider d'un format commun avec nos partenaires européens. Cette réunion permettra de discuter de l'opportunité d'un engagement plus fort d'IRCOM et du GT2 en particulier dans un groupe de travail européen. Cette concertation devrait permettre d'avancer sur la rédaction de bonnes pratiques pour l'usage de la TEI pour décrire nos transcriptions en délimitant un jeu commun de balises dûment exemplifiées.

En parallèle :

▪ **Aide à la finalisation de corpus**

A la suite de l'inventaire des corpus et des besoins réalisé en 2011-2012 (http://ircom.corpus-ir.fr/wiki/doku.php?id=wiki:enquete), le consortium IRCOM a mis en place, en 2013, une aide technique et financière pour la finalisation de corpus en tant que telle. Par *finalisation*, nous entendons toute amélioration de l'état du corpus (documentation, conversion de formats, normalisation, etc., à l'exclusion de la création de ressources ex-nihilo) pouvant faciliter sa réutilisation, et débouchant nécessairement sur le dépôt auprès d'un des centres de ressources (Cocoon ou SLDR). Un appel d'offre (lancé conjointement avec l'Equipex ORTOLANG) a été publié sur les listes de diffusion *IRCOM*, *Corpus-écrits*, *ListeILF*, *Typoling*, et *Parole*. En dehors de 6 projets identifiés à partir de l'inventaire 2012 qui ont été contactés à notre initiative, 14 autres projets ont répondu à l'appel d'offre. Le détail des actions réalisées et en cours de réalisation est accessible à l'adresse : http://ircom.corpus-ir.fr/wiki/doku.php?id=wiki:actions_d_aide_aux_corpus.

Perspectives 2014

▪ **Les corpus** ayant fait l'objet d'une aide financière seront déposés en 2014. L'IRCOM veillera au bon déroulement de cette opération, en fournissant éventuellement une aide technique. Un nouvel appel d'offre pourra être diffusé courant 2014.

▪ L'expérience acquise à travers cette opération de finalisation pourra être mise à profit pour l'**édition d'un guide pratique** (dans le cadre des activités du groupe de travail 6 « hébergement et archivage pérenne des corpus multimodaux »). En effet, en ce qui concerne les corpus traités durant l'exercice 2013, le frein à la finalisation (comprenant le dépôt en vue d'une diffusion) ne résidait pas tant dans des obstacles d'ordre technique que dans le manque d'information claire et unifiée sur les dispositifs existants. Plusieurs corpus étaient archivables en l'état, et l'apport de l'IRCOM s'est limité à une mise en relation et une explication de la procédure, en rapport avec les motivations scientifiques qui justifient l'archivage pérenne et la diffusion.

▪ **Articulation des besoins spécifiques aux corpus oraux et multimodaux avec le cadre plus large des humanités numériques**

Il est apparu, en revanche, que les usagers ont une attente en matière de normalisation de métadonnées métier. Il s'agit d'informations détaillées, pertinentes au regard des objectifs de recherche et par conséquent dont les éléments peuvent varier largement d'une étude à une autre. Ces métadonnées doivent être considérées comme des objets distincts des métadonnées descriptives (ex. OLAC-DC, CMDI, RDFa…) qui accompagnent chaque corpus et/ou ressource multimodale déposée auprès des centres de ressources comparables à des fiches de bibliothèque.

Les métadonnées métiers, au contraire, constituent en elles-mêmes des ressources, en ce sens qu'elles capturent un aspect de la réalité qui est l'histoire du sujet parlant, le contexte de production ou d'interaction, etc. Ces descriptions s'apparentent aux ressources secondaires (transcription ou annotation de l'audio ou de la vidéo), car elles impliquent une part d'interprétation de l'opérateur humain. Il s'agit donc de ressources précieuses qui justifient une prise en charge active en vue de leur conservation et valorisation.

La demande des usagers en matière de normes pour formater les métadonnées métiers traduit leur préoccupation quant au potentiel de réutilisation de ces informations. Il est admis que l'archivage n'a pas comme fin en soi la conservation des ressources, mais constitue plutôt un *moyen* qui permet le développement cumulatif des connaissances. C'est donc l'intelligibilité des métadonnées métier dans le contexte des *humanités numériques* qui est ici questionnée. Derrière cette demande de formalisme, on peut donc déceler une interrogation sur les techniques qui permettront de lire et donc de rendre intelligible le contenu (pertinent dans un cadre de recherche donné) de ce nouveau type d'objet.

Bilan du groupe 2 à mi-parcours

A partir de l'inventaire des corpus réalisés par l'IRCOM, le groupe de travail 'Interopérabilité' a pu mesurer avec précision la diversité des ressources et des formats des corpus oraux et multimodaux de notre communauté.

Dans un premier temps, le groupe s'est focalisé sur les tâches de transcription qui enrichissent les recueils de données et constituent la base des recherches à l'oral. Le groupe s'est réuni par téléconférence et en présentiel pour échanger sur les différentes options disponibles à ce jour et les solutions déjà opérationnelles dans les laboratoires suivant leurs besoins et leurs spécificités. Un premier délivrable est un ensemble de ressources mises en

ligne, documentées et actualisées sur le site web de l'IRCOM pour partager et diffuser les informations à l'ensemble des unités.
Dans un second temps, certains de ses membres engagés dans des instances nationales ou internationales ont contribué à une veille technologique sur les formats standards comme Cimdi ou la Tei pour amorcer le travail de définition d'un format d'échange unifié et normalisé pour faciliter la réutilisabilité des données.

Groupe 4 : Multimodalité et modalité gestuo-visuelle

Objectifs
L'intitulé de ce groupe de travail, « Multimodalité et modalité visuo-gestuelle », souligne un regroupement de problématiques variées qui émergent des recherches concernant trois domaines :

1. l'étude des langues vocales, appréhendées dans l'intégralité de leur contexte multimodal (dont par exemple la gestualité coverbale) ;
2. l'étude des langues des signes, de modalité visuo-gestuelle par nature ;
3. l'étude de la gestualité, en tant que telle, comme une modalité d'expression en soi.

Ces domaines font intervenir des pratiques de recherche variées, autour de corpus de productions multi- ou mono-modales qui sont toutes envisagées comme des productions « orales » (par opposition aux productions « écrites »). Chaque domaine peut être associé à différents types de corpus, mais un même corpus peut relever des trois domaines. Ces corpus impliquent des problématiques communes, ainsi que spécifiques à chaque domaine, concernant leur constitution, dépouillement, codage et analyse, et liées à l'utilisation (conjointement ou non) de différentes modalités d'expression en contexte (gestes, regards, postures, mimiques, autres événements et entités en présence, etc.).

Activités en 2013 et planification pour 2014
Planifications des actions 2013 et 2014 : une réunion s'est tenue en mai 2013. Nous avons planifié l'organisation des formations qui ont eu lieu en novembre 2013, ainsi qu'une proposition d'une journée d'étude consacrée à la « Multimodalité : nouvelles questions, nouvelles méthodes, perspectives interdisciplinaires » dont la thématique permettra d'aborder des réflexions autour des méthodologies nécessaires, par exemple pour inclure des informations variées dans les corpus, parmi lesquelles des données de suivi oculaire, de capture de mouvement, et de type prosodie (tout en élargissant le propos au-delà du système phonologique du français).

Rappel sur le Programme et objectifs des formations
Au cours de la période de labellisation (2011-2015), le programme des formations du groupe 4 vise à balayer l'ensemble de la vie des corpus multimodaux : de la constitution jusqu'à leur dépôt sur des sites d'archivage pérenne et la valorisation de ces corpus en passant par l'annotation et l'analyse des données.
D'une manière générale chaque formation est dispensée deux années consécutives et forment une cinquantaine de participants.

Fin 2013, la première phase —Constitution des corpus— est finalisée, ayant été répétée depuis 2012. Son contenu a donc vocation à constituer un portefeuille de formation.

En effet, le groupe 4, en accord avec les formateurs, décide que les ateliers dispensés deux fois ont vocation à enrichir un portefeuille de formations que proposerait l'IRCOM, mis à disposition des écoles doctorales qui pourraient financer la tenue de sessions pour leurs étudiants, dès lors que leur fréquence ne dépasse pas deux fois l'an.

Fin 2014, la deuxième phase —Annotation des corpus— sera complétée par la formation à un logiciel d'annotation et de requêtes utilisé dans l'acquisition (CLAN) en sus de la formation dispensée en 2013 sur le logiciel ELAN.
L'année 2015 devrait donc s'achever avec une formation renouvelée au logiciel CLAN de deux journées et une formation à l'archivage pérenne et à la valorisation que le groupe 4 pourrait co-organiser avec le groupe 6 créé cette année et chargé de « l'hébergement et de l'archivage pérenne des corpus multimodaux ».

▪ Actions menées — formations
Le groupe 4 a organisé une formation sur la « Constitution, le traitement et l'analyse de corpus multimodaux » au cours de laquelle deux types de médias ont été abordés : l'audio et la vidéo. Cette formation déjà offerte en 2012 se déroule sur deux journées et alterne des présentations concernant les outils et méthodes avec des travaux pratiques de maniement de caméras et d'enregistreurs en situation ainsi que les logiciels utilisés en post-traitement. La présence de trois formateurs a favorisé l'abord de nombreuses thématiques liées à la prise de données, tant techniques (codec, balance des blancs, placement des microphones…) que méthodologiques (choix du cadrage, adéquation du dispositif d'enregistrement aux données primaires à analyser…). Cette formation a été suivie par 22 participants. En 2012 la même formation avait été suivie par 26 participants.

Une formation sur « Notation, annotation et analyse de corpus multimodaux avec ELAN » a réuni 28 participants (17 dans le groupe débutant et 11 dans le groupe avancé) autour de trois formateurs pendant deux journées. La répartition en deux groupes de niveaux a permis de couvrir une grande part des besoins exprimés par les participants.

▪ Actions menées — inventaire et soutien
Afin de compléter l'inventaire des corpus disponibles sur le site de l'IRCOM, une répartition des contacts avec les laboratoires a été faite parmi les membres du groupe 4. Cette action doit être poursuivie et intensifiée en 2014.
Une journée d'étude s'est tenue le 24 mai 2013, *Perspectives croisées sur l'analyse multimodale de corpus oraux*, directement en lien avec les objectifs du groupe 4 et avec le soutien de l'IRCOM. Une synthèse des travaux de cette journée a été envoyée au comité de pilotage de l'IRCOM. Elle figure sur le site.
Le 23 novembre 2013 un colloque lié au projet CREAGEST a exposé les objectifs et la constitution et le début de l'analyse de corpus de LSF. Une synthèse des travaux sera accessible sur le site IRCOM avant la fin de l'année.

Préparation des formations pour 2014
Deux journées de formation consacrées au logiciel CLAN seront organisées à Poitiers à l'automne 2014, pour lesquelles le comité d'organisation est déjà constitué. La formation sur la « Notation, annotation et analyse de corpus multimodaux avec ELAN » sera reconduite pendant deux autres journées, à Poitiers également. Dans les deux cas, le choix du lieu vise à décentraliser les activités de façon à y associer un maximum de participants sur le plan national.

Groupe 5 : Questions juridiques et éthiques, droits des personnes et des producteurs de corpus

Objectifs
Ce groupe de travail s'est donné pour objectif prioritaire la constitution d'une information documentaire de base, à la fois fiable et synthétique, afin d'établir un cadre de traitement global concernant les cas de figure les plus fréquents auxquels sont confrontés les chercheurs en sciences du langage travaillant sur des documents sonores (et audiovisuels) dans leurs pratiques. Dans un premier temps, ont prévalu une approche d'expertise à large spectre et une série de consultations ponctuelles. Après la journée commune organisée avec les consortiums corpus écrits et CAHIER à la fin de l'année 2012, une nouvelle rencontre a été prévue dans une configuration proche au début de l'année 2014. Le consortium Archives des ethnologues pourrait y être associé, s'il le souhaite.

Activités en 2013
Pour assurer à la communauté, de façon pragmatique, les réponses attendues à des interrogations liées à l'enquête et à l'expérimentation, l'IRCOM s'est engagée par voie contractuelle au titre d'une prestation extérieure auprès de deux unités de recherche du CNRS spécialisées dans le droit des biens immatériels, le CECOJI (Centre d'Etudes sur la Coopération Juridique Internationale – laboratoire localisé à Paris et Poitiers, avec une spécialisation concernant les questions de protection de la vie privée) et l'ERCIM (Equipe de Recherche Créations Immatérielles et droit, de l'Université de Montpellier, dans lequel une équipe se consacre à la propriété intellectuelle). Une demande leur a été soumise concernant une mission d'information et de conseil afin que soit défini ce que doivent être les conditions de collecte, de traitement, de conservation et de diffusion des données en conformité avec les cadres juridiques actuels, en tenant compte de la compatibilité du droit français avec le droit européen mais également avec les cadres légaux en usage dans les pays de collecte (une question particulièrement sensible auprès de communautés encore peu impliquées dans le processus de mondialisation, en Afrique subsaharienne, en Asie centrale ou en Amérique latine) et des aires de diffusion où les contraintes et les exigences s'avèrent variables d'une zone à l'autre, notamment en ce qui concerne les accès sur la Toile.

Au cours de l'année 2013, un premier point sur les réalisations a fait l'objet d'un exposé lors de la journée organisée par l'IRCOM en février à laquelle a participé Marie Cornu, directrice du CECOJI. Un rapport d'étape scientifique a ensuite été présenté lors d'une journée construite en concertation entre l'IRCOM, la Bibliothèque nationale de France et les unités de recherche juridiques le 20 juin, sur le site François-Mitterrand. Après une série de

communications synthétiques impliquant l'ensemble des participants institutionnels, un récapitulatif des questions a été établi, affiné par l'échange conduit entre les chercheurs, les conservateurs et les juristes.

Comme il en a été convenu, le rapport CECOJI-ERCIM sera remis en décembre 2013 afin de permettre une programmation concertée des activités en lien avec les groupes de travail correspondants des consortiums Corpus écrits et CAHIER. Le cadre proposé permettra de positionner l'action des consortiums en regard d'autres initiatives conduites parallèlement, telle que celle conduite par des archivistes et ingénieurs à l'intérieur du consortium Archives des ethnologues ou la liste des recommandations éthiques avancées pour le traitement des masses de données qu'a dressée un groupe de travail plus spécialement concerné par les procédures du TAL.

Parallèlement, un certain nombre de réponses ont pu être apportées, en fonction de demandes ponctuelles via le site et les listes de diffusion, en attendant que soient incluses les interrogations concernant le droit à l'image (multimodal) et l'ensemble des questions d'éthique, avec une attention particulière pour les situations de fragilité testimoniale (enquête auprès d'enfants, de personnes handicapées et plus généralement tout public pour lequel le droit à l'information ne peut être assuré dans les conditions prescrites).

Au terme de deux années de fonctionnement, le GT 5 disposera d'une synthèse sans équivalent qui actualisera le *Guide des bonnes pratiques / Corpus oraux* dirigé par O. Baude et qui permettra de faire retour vers l'ensemble de la communauté des chercheurs afin de constituer les éléments d'une jurisprudence à partir des questionnements issus des enquêtes et des archives. La première étape sera la rédaction d'un engagement du témoin, du chercheur et des acteurs du processus pour assurer la protection de leurs droits et un bon usage des ressources, en suivant une orientation qui privilégie les Creative Commons et l'Open source.

Groupe 6 : Hébergement et archivage pérenne des corpus oraux

Objectifs

Le sixième groupe de travail a été lancé en octobre 2013. L'objectif de ce groupe est de répondre aux attentes de la communauté des chercheurs et usagers en matière de pérennisation de corpus oraux et multimodaux. Pour atteindre cet objectif, il se donne pour missions : (i) d'établir une documentation claire et unifiée des solutions existantes en la matière ; (ii) de diffuser et rendre accessible l'information le plus largement possible : réunion, site web, guide des bonnes pratiques, questions/réponses, exemples.

Activités en 2013
Le démarrage du groupe de travail étant récent, les actions 2013 se limitent à une prise de contact avec les principaux interlocuteurs (Huma-Num, ORTOLANG, Cocoon, SLDR, IRCOM) et la programmation d'une première table ronde (prévue pour le 30 janvier 2014 à Paris). Cette réunion vise à établir un premier état des lieux et un cahier des charges de la documentation et des actions à venir.

Perspectives 2014 et 2015
2014 : Deux autres rencontres sont prévues au cours de l'année 2014 (mars-avril et septembre), dont au moins une pourrait porter sur les métadonnées. Le travail mené au cours de l'année 2014 sera restitué à la communauté, en même temps que la documentation, sous forme d'une journée d'étude et d'information, qui se tiendra en fin d'année (novembre). Cette journée sera l'occasion d'un échange et d'une réflexion avec l'intervention d'experts en matière d'hébergement et d'archivage de corpus.

2015 : Une action de large diffusion auprès des usagers, qui sera également l'occasion d'un retour sur le travail effectué par le GT6, est prévue en 2015 sous la forme de deux journées de formation à l'archivage de corpus (avec démonstration 'on-line' d'un dépôt de corpus dans un Centre de ressources à la clé).

4 - ANNEXE 1 : LE WIKI ET LE SITE WEB

Le wiki : http://www.ircom.corpus-ir.fr/wiki/
Il s'agit d'un espace privé permettant d'agréger toutes les informations rassemblées par les membres du consortium et les groupes de travail.
Les informations du wiki :
- des informations relatives à l'IRCOM (dates, objectifs, membres, etc.)
- une description et une présentation des 5 groupes de travail
- un glossaire
- un annuaire des laboratoires travaillant sur des corpus oraux et multimodaux au niveau national, leurs projets et leurs besoins
- un inventaire détaillé des corpus oraux et multimodaux : nature des données, état et disponibilité
- des ressources méthodologiques ou logicielles
- différentes actualités (ateliers, formations, appels d'offres, etc.)

Les données inventoriées et décrites :
- 50 laboratoires
- 75 corpus

Les ressources :
- Glossaire, manuels d'utilisation, comparatif des logiciels d'annotation, outils de conversion

Passage du wiki au site web :

Principes :

- une seule source d'informations : le wiki
- un site web avec des informations validées
- des fonctions de tri pour exploiter les données comme les corpus ou les laboratoires

Modalités de mises à jour (documentation en ligne)

- aucune information ne peut être ajoutée directement sur le site
- toutes les mises à jour sont faites à partir des données présentes sur le Wiki
- un mode administrateur sur le site web qui "importe" les données actualisées (ajout, modification) du wiki et affiche les nouvelles pages

Contraintes

- page-type pour le parsing

Le site web : http://ircom.corpus-ir.fr/site/accueil.php

Il s'agit d'un espace public permettant la diffusion d'informations à l'ensemble de la communauté : informations, données et ressources du wiki (cf. supra) dont glossaire, mais aussi des listes de diffusion permettant un échange direct par mail entre tous les membres du consortium ou entre les membres d'un groupe de travail.

5 - Annexe 2 : Rappel de la composition du Comité de pilotage

Le Comité de pilotage s'est doté d'une présidence tournante (pour une durée d'un an). Certains membres ont demandé à être remplacés.
Voici la composition actuelle :

1. Stéphane Robert (Fédération TUL) – **Porteur du consortium**
2. Amina Mettouchi (LLACAN) - *Présidence octobre 2011-12*
3. Christophe Parisse (Modyco) - *Présidence octobre 2012-13*
4. Harriet Jisa (DDL) – *Co-Présidence 2012-13*
5. Gabriel Bergounioux (LLL)
6. Carole Etienne (ICAR)
7. Martine Adda **>** Cedric Gendrot (LPP)
8. Boyd Michailovsky **>** Evangelia Adamou (LACITO)
9. Philippe Blache **>** Roxane Bertrand (LPL)
10. Maya Hickmann **>** Dominique Boutet (SFL) - ***Présidence octobre 2013-14***
11. Catherine Bolly (SFL) (intégration octobre 2013)