Les groupes de travail
du consortium corpus oraux et multimodaux
IRCOM

Les groupes suivants ont été constitués à la suite d'une réflexion du Comité de Pilotage et du Conseil Scientifique de l'IRCOM.
Vous pouvez les rejoindre si vous souhaitez participer à la réflexion, en contactant les coordinateurs, dont les mails vous sont donnés en début de descriptif pour chaque groupe.

Le résultat des travaux de ces groupes sera mis en ligne sur le site public de l'IRCOM (en préparation) courant 2012.
En attendant, les informations concernant les travaux du Consortium sont accessibles sur la page de la fédération TUL :
http://www.typologie.cnrs.fr/spip.php?rubrique5&lang=fr

Objectifs généraux des groupes de travail :

Les groupes de travail ont pour objectif de contribuer à la mise en œuvre des différentes missions de l'IRCOM en se focalisant sur une problématique spécifique, souvent en relation avec celles d'autres groupes de travail (le découpage n'étant pas orthogonal). Ces différentes missions visent à :

- organiser et accompagner le développement de corpus oraux et multimodaux en linguistique
- aider les chercheurs à s'approprier les outils nécessaires et à développer des standards communs de référence
- aider à développer la visibilité, l'accessibilité et la valorisation des fonds existants
- aider à améliorer leur mise à disposition, leur mutualisation et leur interopérabilité afin d'intégrer les réseaux internationaux (notamment ERIC-CLARIN)
- intégrer la communauté des producteurs et utilisateurs de corpus oraux et multimodaux dans ces pratiques et réflexions.
- faire le point sur l'état des actions, corpus, recherches et travaux en France et au niveau international par rapport à la thématique du groupe de travail.

Groupe de travail n°1

Corpus oraux et multimodaux : finalités scientifiques

Coordinateurs : Philippe Blache, Amina Mettouchi, Laurence Devillers, Anne Lacheret-Dujour, Sophie Rosset
Participants : Paul Cappeau, Brigitte Garcia, Philippe Boula de Mareuil, Dominique Boutet, Annelies Braffort, Cyril Auran, Martine Adda-Decker, Laurent Gautier, José Deulofeu, Tobias Scheer, (à compléter…).

Contacts :
Philippe Blache : philippe.blache@lpl-aix.fr
Laurence Devillers : laurence.devillers@paris-sorbonne.fr
Anne Lacheret-Dujour : anne@lacheret.com
Amina Mettouchi : mettouchi@vjf.cnrs.fr
Sophie Rosset : rosset@limsi.fr

L'objectif de ce groupe de travail « Finalités Scientifiques » 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.

La première question à laquelle notre groupe sera amené à réfléchir est celle de la définition de notre objet : qu'est-ce qu'un corpus de linguistique ?
De ce point de vue, quelles sont les relations entre les corpus issus de la "linguistique de laboratoire", de la "linguistique de terrain", de la "linguistique expérimentale", de la "linguistique théorique", de la "linguistique appliquée" (notamment en didactique des langues ou dans le domaine des discours spécialisés/professionnels) dans un contexte où toutes peuvent avoir recours à des ressources constituées en corpus ?
Plus précisément, y-a-t-il des possibilités de transferts méthodologiques entre les corpus appartenant à différents champs des sciences du langage (terrain, TAL (parole), psycholinguistique, pragmatique, etc.), pouvant mener à des collaborations interdisciplinaires ?
Egalement, comment se définissent les corpus oraux et multimodaux, en contraste et continuité avec les corpus écrits, ainsi que ceux issus du web ?

Qu'entend-on par données primaires? Données secondaires ? En lien avec ce double questionnement, qu'entend-on respectivement par "transcription" et "annotation" (cette question sera également abordée par le groupe 4) ?
Qu'est-ce exactement que la "modalité" et la "multimodalité" (au sens, entre autres, où les données "gestuelles" peuvent, aussi, être considérées comme relevant de plusieurs modalités, visuelle/gestuelle/proprioceptive) ?

En règle générale, un corpus est constitué en fonction des finalités de recherche fondamentale ou appliquée auxquelles il est destiné ce qui influe sur un certain nombre de choix d'annotation, entre autres.
Une des questions que l'on peut alors se poser est celle de la réutilisation possible de corpus déjà annotés à des fins de recherche différentes. Cette réutilisation est-elle possible ? Sous quelles conditions ? Par exemple, quel type d'information et quel degré de précison sont nécessaires pour une exploitation diversifiée des métadonnées ?
Cette question devra être traitée en complémentarité avec le groupe de travail « Interopérabilité », qui se focalise sur les solutions logicielles et les standards. Elle sera vraisemblablement déclinée en fonction des champs de recherche.
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). Ce travail pourra être mené en collaboration avec le groupe « corpus multilingues ».
Dans le même ordre d'idée, peut-on considérer qu'il existe un corpus « minimal » (comprenant certains types de métadonnées, et certains types d'annotations, que l'on pourrait considérer comme « de base ») ?

Notre groupe a prévu également de proposer un bilan des corpus qui ont permis des progrès importants dans le champ de la linguistique, et identifier par conséquent des bonnes pratiques.
On pourra parallèlement s'interroger sur ce qui manque le plus en matière de corpus oraux/multimodaux (entre autres pour une langue telle que le français (variation diastratique par exemple)).
Dans cette même perspective rétrospective, peut-on constater un changement dans les catégories d'analyse suite à l'utilisation de l'informatique pour le traitement des données ?

Sur le plan méthodologique, notre groupe sera enfin amené à réfléchir aux étapes de constitution d'un corpus, « un corpus de A à Z », en fonction des réflexions que nous aurons menées.

Modalités de fonctionnement

Le groupe fonctionnera par divers types d'échanges réguliers (forum Wiki, courriels, téléconférences, réunions si nécessaire) et devra fournir des comptes rendus au Comité de Pilotage IRCOM plusieurs fois par an. Il est envisagé d'organiser des rencontres avec d'autres groupes de travail.

Groupe de travail n°2

Interopérabilité

Coordinateurs : Christophe Parisse et Carole Etienne
Participants : Nicolas Ballier, Bernard Bel, Christophe Benzitoun, Philippe Blache, Annelies Braffort, Dominique Fohr, Clément Planck (à compléter)

Contacts :
Carole Etienne : carole.etienne@ens-lyon.fr
Christophe Parisse : cparisse@u-paris10.fr

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 s'heurte donc 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 d'apporter des solutions à cette problématique qui se reposent sur les outils existants et qui s'ouvrent sur une amélioration des logiciels, des corpus, et des outils et formats d'échange entre corpus.

L'inventaire en cours de réalisation à l'IRCOM permettra d'identifier les corpus oraux et multimodaux dans nos laboratoires et leurs formats mais aussi les pratiques qui les accompagnent pour décrire ces ressources ou enrichir les données par des transcriptions et des annotations. A partir de ce recueil, le groupe « Interopérabilité » aura une meilleure connaissance des métadonnées (descripteurs) et des logiciels utilisés ainsi que des 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 laboratoires.

En parallèle, il connaîtra les différentes solutions mises en œuvre dans les équipes et pourra coordonner les réflexions pour mutualiser ces solutions et les pérenniser dans la communauté.

Le groupe de travail portera ses efforts sur deux axes :

- les standards de données et les normes, au niveau national et international, pour garantir la pérennité de nos données orales et démystifier des standards difficiles à appréhender au niveau d'un projet, d'une équipe de recherche ou d'un laboratoire.
- les logiciels de transcription et d'annotations dont les formats souvent propriétaires rendent difficiles le passage des données d'un logiciel à l'autre, d'un environnement à l'autre (mac, pc, linux, ...) tâches pourtant nécessaires quand aucun des logiciels ne remplit à lui seul toutes les fonctions dont nous avons besoin pour nos travaux.

Le statut fédérateur du groupe de travail permettra d'engager des 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 ces standards d'une manière satisfaisante pour notre communauté.
De même, ce statut facilitera les négociations avec les concepteurs des logiciels pour améliorer les fonctions d'export et éventuellement d'autres fonctions pour faciliter leur exploitation dans nos unités.
Le groupe de travail informera la communauté sur le suivi des demandes d'évolutions pour les normes comme pour les logiciels.

Le groupe de travail aura pour mission d'alimenter le guide en ligne des bonnes pratiques pour:

- diffuser des informations sur les logiciels et sur les normes, leur stabilité et leur interopérabilité,
- mettre en ligne les tutoriaux des logiciels,
- délivrer des exemples explicites de corpus oraux dans les principaux standards (Dublin Core, Tei, Clarin ou Dariah, ...) interopérables avec un ensemble de logiciels massivement utilisés dans la communauté,
- donner des exemples de fonctions réalisables dans les principaux logiciels avec leurs formats d'entrée et de sortie,
- proposer 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.

Groupe de travail n°3

Corpus multilingues

Coordinateurs **:** Martine Adda-Decker, Boyd Michaïlovsky, Stéphane Robert, Pascal Nocera, Philippe Boula de Mareüil
Participants : Isabelle Léglise, Gudrun Ledegen, Valentina Vapnarsky, Véronique Traverso, Amina Mettouchi, Brigitte Garcia (à compléter)

Contacts :
Martine Adda-Decker : madda@limsi.fr
Philippe Boula de Mareuil : Philippe.Boula.de.Mareuil@limsi.fr
Boyd Michaïlovsky : boydm@vjf.cnrs.fr
Pascal Nocera : pascal.nocera@univ-avignon.fr
Stéphane Robert : robert@vjf.cnrs.fr

1. Objectifs et périmètre

Le groupe de travail « Corpus multilingues» regroupe des problématiques variées concernant :

1) l'étude de langues ou variétés de langues « rares », sans tradition écrite, qui ne sont pas ou que peu décrites; les variétés de langues d'apprenants, de locuteurs bilingues et plurilingues ; les créoles, les langues en contact ; le code-switching…
2) la production de corpus incluant plusieurs langues, soit en parallèle ou en simultané au niveau source (audio), soit au niveau des annotations ; la création de corpus parallèles avec des contenus similaires dans différentes langues; l'annotation multilingue de corpus de langue des signes…
3) le développement de méthodes et d'outils « indépendants de la langue », ou au moins utilisables pour un nombre élevé de langues. Cette thématique est en lien avec le groupe interopérabilité.

Sont concernés par ces thématiques les chercheurs ayant une expertise de recherche dans différentes langues, une expertise dans l'adaptation de leurs méthodes de recherche d'une langue à l'autre ou une expertise dans le domaine du plurilinguisme et du contact de langues. De même, sont concernés des chercheurs devant produire des annotations/gloses en différentes langues et des chercheurs en traitement automatique de la parole (transcription, traduction, synthèse, recherche d'information…), pour lesquels le déploiement de technologies existantes pour un grand nombre de langues est un défi majeur. Ce déploiement génère souvent des questions de recherche d'ordre linguistique et s'accompagne généralement de la création de grands corpus multilingues. Les domaines d'application de notre groupe de travail incluent les nouveaux usages pour l'apprentissage des langues et les technologies vocales multilingues ainsi que les problèmes qui se posent pour l'annotation de corpus de langues pas ou peu décrites.

2. Actions envisagées

- Recenser et regrouper les unités, équipes et individus concernés.

- Identifier et centraliser les besoins ainsi que les ressources disponibles : corpus, tutoriels, métadonnées, manuels de codage, guide de bonnes pratiques…

- Établir un état des lieux en identifiant les problèmes généraux qui traversent les différentes problématiques visées par le groupe de travail ainsi que les spécificités de chacune d'entre elles, en impulsant de nouvelles pratiques/questions de recherche, notamment à l'interface entre domaines.

- Augmenter l'utilisation et la visibilité des ressources multilingues à travers les différentes équipes (et disciplines) scientifiques concernées, mais aussi au niveau international.

- Promouvoir la standardisation des corpus et en améliorer la spécification en lien avec les autres groupes de travail concernés.

- Identifier les besoins en formation

- Recenser/lancer des initiatives francophones/européennes/internationales de création de corpus multilingues.

- Faire le point sur l'état des travaux (recherches, création de corpus, normes, outils…) au niveau international dans les domaines de recherche concernés par les corpus multilingues ou plurilingues.

3. Modalités de fonctionnement du groupe de travail

Le groupe fonctionnera par divers types d'échanges réguliers (courriels, téléconférences, réunions si nécessaire) et devra fournir des comptes rendus au Comité de Pilotage IRCOM plusieurs fois par an.

Il pourra faire appel à des experts externes nationaux ou internationaux pour éclairer ses travaux.

Il est envisagé également d'organiser des rencontres avec d'autres groupes de travail, notamment dans les divers consortiums (Corpus Oraux et Multimodaux, Corpus Ecrits).

Groupe de travail n°4

Multimodalité et modalité visuo-gestuelle

Coordinateurs: Maya Hickmann, Harriet Jisa, Dominique Boutet & Brigitte Garcia
Participants : Participants : Annelies Braffort, Jean-Marc Colletta, Christophe Parisse, Valentina Vapnarsky, Gaelle Ferré… (à compléter)

Contacts :
Dominique Boutet : dominique_jean.boutet@orange.fr (lire : dominique_jean.boutet)
Brigitte Garcia : bridge.garcia@wanadoo.fr
Maya Hickmann : maya.hickmann@sfl.cnrs.fr
Harriet Jisa : Harriet.Jisa@univ-lyon2.fr

1. Introduction : intitulé, domaines, périmètre

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 co-verbale) ;

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.).

2. Actions envisagées

- Recenser et regrouper les unités, équipes et individus concernés.

- Identifier et centraliser les besoins ainsi que les ressources disponibles : corpus, tutoriels, métadonnées, normes de diffusion, manuels de codage, pratiques, etc.

- Etablir un état des lieux en identifiant les problèmes généraux qui traversent les trois domaines ainsi que les spécificités de chacun d'entre eux et en impulsant de nouvelles questions de recherche, notamment à l'interface entre domaines.

- Optimiser l'utilisation et la visibilité des ressources, par exemple : élaborer un « guide de bonnes pratiques », examiner différentes modalités de diffusion, faire les aménagements nécessaires pour accroître leur accessibilité, y compris par des moyens de visualisation pour les articles scientifiques.

- Promouvoir la standardisation des corpus et en améliorer la spécification, par exemple : faire le point sur différents systèmes et critères de segmentation automatique ou semi-automatique, mettre en place une police de caractères pour la notation des gestes, proposer des avancées dans l'analyse de la langue des signes au vu des limites des gloses en langue vocale ainsi qu'en matière de bases de données lexicales (lemmatisation) et de procédures de transcription, recenser les développements d'outils paramétrables permettant de segmenter le signal, etc.

- Recenser les besoins en formation, notamment en ce qui concerne : les enregistrements lors de la constitution des corpus ; le découpage, l'annotation et l'analyse des données dans différents systèmes ; ainsi que la diffusion des métadonnées et le dépôt d'informations sur un centre de ressources.

3. Constitution et fonctionnement du groupe de travail

- Participants au groupe (liste ouverte à compléter)
  Lors de la réunion du Conseil Scientifique qui s'est tenue le 16 janvier 2012, les personnes ci-dessous se sont inscrites pour participer à ce groupe de travail.

Prénom	**Nom**	**Rôle**	**Laboratoire**
Maya	Hickmann	C.P. co-pilote	Laboratoire Structures Formelles du Langage, CNRS & Paris 8
Harriet	Jisa	C.P. co-pilote	Laboratoire Dynamique du Langage, CNRS & Lyon 2
Dominique	Boutet	C.S. co-pilote	Laboratoire Structures Formelles du Langage, CNRS & Paris 8
Brigitte	Garcia	C.S. co-pilote	Laboratoire Structures Formelles du Langage, CNRS & Paris 8
Annelies	Braffort	CS	LIMSI-CNRS, Orsay
Jean-Marc	Colletta	CS	Laboratoire LIDILEM, Grenoble
Gaëlle	Ferré	CS	LLING, Laboratoire de Linguistique de Nantes
Aliyah	Morgenstern	CS	Langues, Textes, Arts et Cultures du Monde - Sorbonne Nouvelle, Paris 3
Christophe	Parisse	CP	Laboratoire Modyco, CNRS & Paris 10
Valentina	Vapnarsky	CS	CNRS-Centre EREA- Laboratoire d'Ethnologie et de Sociologie Comparative

- Experts externes
  Le groupe identifiera et pourra solliciter des experts externes nationaux ou internationaux.

- Modalités de fonctionnement
  Le groupe fonctionnera par divers types d'échanges réguliers (courriels, téléconférences, réunions si nécessaire) et devra fournir des comptes rendus au Comité de Pilotage IRCOM plusieurs fois par an. Il est envisagé d'organiser des rencontres avec d'autres groupes de travail.

Groupe de travail n°5

Questions juridiques et éthiques, droits des personnes et des producteurs de corpus

Coordinateurs: Gabriel Bergounioux (Comité de Pilotage / LLL – Orléans), Bernard Bel (Conseil scientifique / LPL – Aix-Marseille) et Sophie Rosset (Conseil scientifique / LIMSI – Paris-Sud)
Participants : Laurence Devillers (Comité scientifique / LIMSI – Paris-Sud), Cécile Fougeron (Comité scientifique / LPP – Paris 3), Emilie Jouin (ICAR – ENS Lyon), Aliyah Morgenstern (Comité scientifique / PRISMES – Paris 3), Claire Moyse-Faurie (LACITO), Marie-Anne Sallandre (Comité scientifique / SFL – Paris 8) (à compléter)

Contacts :
Bernard Bel : bernard.bel@lpl-aix.fr
Gabriel Bergounioux : gabriel.bergounioux@univ-orleans.fr
Sophie Rosset : rosset@limsi.fr

Le groupe de travail réfléchit sur toutes les questions juridiques et éthiques qui ont des implications ou des incidences en matière de corpus oraux et multimodaux D'un côté, il se propose de recenser, réunir et évaluer l'ensemble des acquis et des expériences accumulées dans les laboratoires et les équipes, de centraliser l'ensemble des initiatives et des thèses qui ont participé à la définition du champ et de ses enjeux et d'établir la liste des questions que se posent les chercheurs, des réponses qu'ils ont pu y apporter. De l'autre, ce groupe entend aider au conseil et à la prescription dans le domaine, à toutes les échelles, y compris internationale.

Au nombre des points qui seront traités :

Les ayants droit

- témoins et participants extérieurs présents sur l'enregistrement
- enquêteur (et son institution) et assistants de production éventuellement
- transcripteurs et annotateurs (définition des contrats de travail)
- propriétaire des logiciels (recommandation pour les Creative Commons)
- maquettistes et producteurs de métadonnées
- services de versement, d'hébergement, de maintenance, de mise à disposition
- sources de financement
- usages commerciaux et droits dérivés

Le droit de propriété

- nature de la responsabilité juridique (et pénale)
- protection des auteurs (cf. loi sur le patrimoine)
- établissement d'un copyright et application du droit d'auteur
- formalité de dépôt en archive (procédure de dépôt légal ?)
- statut des corpus oraux recueillis sur Internet

L’aliénabilité

- signification de l’aliénabilité pour un corpus oral ou multimodal
- droit à l’utilisation (privée, institutionnelle, publique…)
- contraintes concernant la consultation
- fractionnement de l’aliénabilité (prosodie, transcriptions, métadonnées…)

La protection juridique et éthique

- la définition de l’anonymisation
- les recours contre les producteurs de documents sonores
- les protocoles d’enquête et les publics sensibles
- le retour aux producteurs de leurs productions
- le droit à l’image des communautés de population

Le droit international

- les différences de juridiction selon les Etats
- la spécificité de législation sur Internet
- les coopérations à l’échelle européenne (CLARIN, DARIAH…) et au-delà

Les questions d’éthique

- droit moral
- protection des publics sensibles

*****