 

**CONSORTIUM CORPUS ORAUX ET MULTIMODAUX**

**Appel IRCOM-ORTOLANG 2014**

**Aide technique et financière à la finalisation de corpus**

Le consortium IRCOM et l’Equipex Ortolang ont lancé conjointement en juin 2013 un appel pour l’aide à la finalisation de corpus.

Vous avez été nombreux à répondre à cet appel. Au total, 20 projets ont été retenus et ont fait, ou feront l’objet, d’une aide technique, parmi lesquels 9 projets ont bénéficié, ou bénéficieront d’une aide financière (<http://ircom.huma-num.fr/site/p.php?p=corpusaide>).

**En 2014,** **le consortium IRCOM conjointement avec l’équipex Ortolang proposent**

**de poursuivre son accompagnement technique et financier.**

L’accompagnement technique a comme objectif d’amener les ressources linguistiques, comportant de préférence des annotations (ex. transcriptions), à être :

(1) stockées de manière sécurisée,

(2) référencées par un identifiant unique international

(3) à être diffusée (publiquement ou sous conditions, immédiatement ou avec délai).

Cette aide peut intervenir concrètement à plusieurs niveaux : conversion de format des fichiers, assistance à la documentation (métadonnées compatibles OAI-PMH), mise en relation avec les différents interlocuteurs en fonction de la spécificité des ressources (centres de ressources/BnF).

Vous trouverez des descriptions de corpus finalisés sur le site IRCOM (<http://ircom.huma-num.fr/site/p.php?p=corpusaide>).

A titre d’illustration, voici quelques corpus qui ont pu être finalisés après avoir bénéficié de notre aide technique (nom du correspondant entre parenthèses).

* Le corpus OTG (Jean-Yves Antoine), composé d’interactions d’une durée totale de 2h a bénéficié de notre aide pour la conversion des 315 fichiers au format brut (.raw) vers le format .aiff, un format accepté pour l’archivage numérique auprès du CINES (qui prend en charge la préservation à long terme des données de la recherche). Cet ensemble a été déposé et est diffusé par SLDR/Ortolang (http://sldr.org/sldr000831).
* L’Atlas Linguistique de l’Alsace (Claude Evrard), composé de trois principaux ensembles d’enregistrements réalisés dans les années 60, 80 et 90 (entretiens ethnographiques, atlas linguistique, enquête sur la conscience linguistique) d’un total de 864 ressources audio, et dont les producteurs ne sont plus en activité, a été documenté en recoupant l’information présente sur le site web (http://ala.u-strasbg.fr/index3.php) et plusieurs fichiers excel conservés. Cet atlas a été déposé auprès de CoCoON (http://cocoon.huma-num.fr/exist/crdo/collection\_ala.htm).
* Le corpus ALIPE (Thierry Chanier), déjà structuré et consultable via un site web universitaire (http://lrl-diffusion.univ-bpclermont.fr/alipe/), a été orienté vers SLDR/Ortolang pour être référencé et stocké de manière sécurisé, et diffusé au format HD (http://sldr.org/alipe-000853). Une attention particulière a été fournie pour que l’arborescence soit conservée, en particulier dans les liens de la documentation (html) vers les fichiers son.
* Le corpus VN-Attitude (Dang Khoa Mac), composé d’audio, de vidéo et de signaux électro-glottographiques (EGG), nécessitait une conversion de format de ces derniers pour qu’ils puissent être réellement interopérables (dépôt en cours auprès de CoCoON). En revanche, l’EGG n’étant pas un type de données pris en charge par le CINES, une réflexion est en cours concernant son statut en tant qu’objet archivistique.

**Conditions de l’appel 2014 d’IRCOM et d’Ortolang à accompagnement technique et financier pour la diffusion de corpus de langage**

Pour palier la grande disparité dans les niveaux de compétences informatiques des personnes et groupes de travail produisant des corpus, L’IRCOM propose une aide personnalisée à la finalisation de corpus. Celle-ci sera réalisée par un ingénieur IRCOM en fonction des demandes formulées et adaptées aux types de besoin, qu’ils soient techniques ou financiers.

Les demandes peuvent concerner tout type de traitement : traitements de corpus quasi-finalisés (conversion, anonymisation), alignement de corpus déjà transcrits, conversion depuis des formats « traitement de textes », digitalisation de support ancien.

* Les demandes sont à déposer du 10 septembre 2014 au 31 octobre 2014.
* Les demandes seront traitées dans l’ordre où elles seront reçues par l’IRCOM.
* Les demandes émanant d’EA ou de petites équipes ne disposant pas de support technique « corpus » seront traitées prioritairement.
* La décision pour l'aide technique relèvera du comité de pilotage d’IRCOM.
* Le montant maximal par projet est de 3300 euros.
* Les demandes non traitées en 2014 sont susceptibles de l’être en 2015.
* Si vous avez des doutes quant à l’éligibilité de votre projet, n’hésitez pas à nous contacter pour que nous puissions étudier votre demande et adapter nos offres futures, vous pouvez contacter Martine Toda [ martine.toda@cnrs.fr ] ou à [ [ircom.appel.corpus@gmail.com](mailto:ircom.appel.corpus@gmail.com) ]
* Cet appel d’offre est soumis aux conditions formulées ci-dessous.

Les conditions nécessaires pour proposer un corpus à finaliser sont :

* Pouvoir prendre toutes décisions concernant l’utilisation et la diffusion du corpus (propriété intellectuelle en particulier).
* Disposer de toutes les informations concernant les sources des corpus et le consentement des personnes enregistrées ou filmées.
* Accorder un droit d’utilisation libre des données ou au minimum un accès libre pour la recherche scientifique.

Les propositions de réponse à cet appel d’offre sont à envoyer avant le 31 octobre 2014 à [ircom.appel.corpus@gmail.com](mailto:ircom.appel.corpus@gmail.com) et doivent utiliser le formulaire de la page suivante. Dans tous les cas, l’IRCOM vous adressera une réponse personnalisée.

Ces propositions doivent présenter les corpus proposés, les données sur les droits d’utilisation et de propriétés et sur la nature des formats ou support utilisés.

Cet appel est organisé sous la responsabilité de l’IRCOM avec la participation conjointe de l’IRCOM (<http://ircom.huma-num.fr>) et de l’EquipEx ORTOLANG (<http://www.ortolang.fr>).

Pour toute information complémentaire, nous rappelons que le site web de l’IRCOM (<http://ircom.huma-num.fr>) est ouvert et propose des ressources à la communauté : glossaire, inventaire des unités et des corpus, ressources logicielles (tutoriaux, comparatifs, outils de conversion), activités des groupes de travail, actualités des formations, ...

L'IRCOM invite les unités à inventorier leur corpus oraux et multimodaux - 70 projets déjà recensés - pour avoir une meilleure visibilité des ressources déjà disponibles même si elles ne sont pas toutes finalisées.

*Le comité de pilotage IRCOM*

Le Consortium IRCOM appartient au TGIR HUMA-NUM - <http://www.huma-num.fr/>

**Utiliser ce formulaire pour répondre à l’appel : Merci.**

**Réponse à l’appel 2014 pour la finalisation de corpus oral ou multimodal**

Nom du corpus :

Nom de la personne à contacter :

Adresse email :

Numéro de téléphone :

Nature des données de corpus :

Existe-t-il des enregistrements :

Quel média ? Audio, vidéo, autre…

Quelle est la longueur totale des enregistrements ? Nombre de cassettes, nombre d’heures, etc.

Quel type de support ?

Quel format (si connu) ?

Existe-t-il des transcriptions :

Quel format ? (papier, traitement de texte, logiciel de transcription)

Quelle quantité (en heures, nombre de mots, ou nombre de transcriptions) ?

Disposez vous de métadonnées (en particulier de l’information concernant les droits d’auteurs et d’usage) ?

Disposez-vous d’une description précise des personnes enregistrées ?

Disposez-vous d’une attestation de consentement éclairé pour les personnes ayant été enregistrées ? En quelle année (environ) les enregistrements ont eu lieu ?

Quelle est la langue des enregistrements ?

Le corpus comprend-il des enregistrements d’enfants ou de personnes ayant un trouble du langage ou une pathologie ?

Si oui, de quelle population s’agit-il ?

Dans un souci d’efficacité et pour vous conseiller dans les meilleurs délais, il nous faut disposer d’exemples des transcriptions ou des enregistrements en votre possession. Nous vous contacterons à ce sujet, mais vous pouvez d’ores et déjà nous adresser par courrier électronique un exemple des données dont vous disposez (transcriptions, métadonnées, adresse de page web contenant les enregistrements).

Nous vous remercions par avance de l’intérêt que vous porterez à notre proposition.

**Pour toutes informations complémentaires veuillez contacter Martine Toda**

**martine.toda@cnrs.fr ou** [ircom.appel.corpus@gmail.com](mailto:ircom.appel.corpus@gmail.com)**.**