Data Management Plan pour les projets du Centre scientifique de compétence sur le plurilinguisme (CSP)

Date : 07.09.2015

Nom du chef de projet	Amélia Lambelet
Titre du projet	Corpus de productions écrites d'enfants issus de la migration
Domaine de recherche	☒ Plurilinguisme individuel ☐ Enseignement et apprentissage des langues, évaluation des compétences langagières ☐ Plurilinguisme institutionnel et sociétal ☐ Autre :
Type de projet	☒ Centre scientifique de compétence sur le plurilinguisme ☐ Institut de plurilinguisme ☐ Mandat ☐ Autre :
Chercheurs participants externes, institutions partenaires et autres mandants	*Nom, prénom (institution, rôle de l'institution)*
Durée estimée du projet	Début : 01-01-2016 Fin : 21-12-2017

Description du projet

Ce projet vise à décrire le développement des compétences productives à l'écrit d'enfants issus de la migration portugaise en Suisse (en langue d'origine et langue de scolarisation). Il se base sur les données récoltées dans le cadre du projet LCO du programme de travail 2011-2015 du CSP. Le but du projet est triple et se déroulera en trois étapes :

1. Constitution d'un corpus de productions écrites en portugais/allemand/français d'élèves de 5ème et 6ème HarmoS.

2. Calcule de diverses mesures de la richesse lexicale de ces productions

3. Etude des transferts de surface apparaissant dans ces productions et autres sous-projets sur la base du corpus constitué.

[…]

Types de données

Couverture géographique

Suisse allemande, Suisse romande et Portugal

Méthode(s) de sélection et procédures de collecte

a) Le corpus a été récolté dans le cadre du projet LCO 2011-2015. Il s'agit de productions écrites en portugais, en allemand et en français : 1) scannées, 2) transcrites ; 3) corrigées (orthographe et grammaire) 4) lemmatisées ; 5) annotées (au niveau du lexique)

b) La diversité et la sophistication lexicale de chaque groupe seront calculées. La diversité lexicale individuelle sera calculée pour chaque enfant […]. La sophistication lexicale sera calculée :
   - de manière intrinsèque avec d'autres corpus de référence
   - et extrinsèque (jugement d'expert): distribution de questionnaires

c) Pour les transferts et d'autres sous-projets, le corpus sera annoté en conséquence.

Approche de recherche *(Si l'approche est mixte cocher les deux cases)*

☒ quantitative ☒ qualitative

Selon vos prévisions, quelles seront les données produites dans le cadre de ce projet ?

☒ Enregistrements statistiques
☒ Transcriptions
☐ Données multimédia (audio, vidéo)
☒ Numérisations
☐ Autre :
☒ Exploitations statistiques
☐ Outils de collecte de données
☒ Fichier texte
☐ Livre-code (table de correspondance pour coder, décoder, décrire, documenter les données)

Dans quels formats les données seront probablement stockées ?
(Veuillez choisir toutes les réponses qui vous conviennent)

☐ Papier imprimé
☒ Document écrit à la main (ex. questionnaires)
☒ Word (.DOC/.DOCX)
☒ Texte PDF (.PDF)
☐ Autres formats de texte
☐ .JPEG/.GIF/.TIFF/.PNG
☒ Image PDF (.PDF)
☐ Autres formats d'image
☒ Excel (.XSL)
☒ Comma-separated values (.CSV)
☐ Extensible Markup (.XML)
☐ .SPSS
☒ .R
☒ Max QDA
☐ Autres formats de feuille de calcul
☐ .OGG
☐ .AVI
☐ .MOV
☒ .MP4
☐ .MP3
☐ .WAV
☐ Autres formats audiovisuels
☐ Autre :

Quels logiciels seront utilisés pour le stockage, le traitement et l'analyse des données ?

☐ SPSS ☒ R ☒ Excel ☒ MAXQDA ☒ Autre : TMX (pas encore certain)

Est-ce que des données déjà existantes d'autres projets (internes ou externes) seront inclus ?

☒ oui ☐ non

Si oui, lesquels ?

- Projet LCO du programme de travail 2011-2015 du CSP (interne)
- Données de corpus représentatifs des trois langues déjà existants

Quelles publications sur ces données et sur leurs résultats sont prévues ?

Article scientifique et participation à une conférence

Aspects éthiques et droit d'auteur

Quels droits accordez-vous au Centre scientifique de compétence sur le plurilinguisme (CSP) sur les données produites ? Quels droits sont-ils éventuellement aux partenaires externes ?

Cliquez ici pour saisir du texte.

Concernant la collecte et la réutilisation des données, quels sont-ils les accords conclus avec les participants dans le cadre de ce projet ?

Je ne peux pas encore compléter cela

Les données permettent-elles de tirer des conclusions sur l'identité des participants à l'étude individuelle ? En particulier, de tirer des conclusions sur l'état de santé, les convictions politiques ou religieuses, l'appartenance syndicale, la sexualité, la situation financière ou l'origine ethnique ?

☒ oui ☐ non

Si oui, ces données seront rendues anonymes ou pseudonymes au cours du projet ?

☒ oui ☐ non

Détails : *Cliquez ici pour saisir du texte.*

Accessibilité des données

Une fois le projet terminé, qui est autorisé à avoir accès aux données ?

☐ tout le monde
☐ les chercheurs de l'Institut du plurilinguisme
☐ seulement les collaborateurs du projet
☒ il est compliqué *(Précisez ci-dessous)*

Certaines informations (les données brutes du corpus) seront accessibles à chacun, tandis que pour d'autres données (ex. certaines métadonnées) il faudra demander une autorisation au chef de projet.

Qui pourrait être intéressé par ces données ? *(Veuillez choisir toutes les réponses qui vous conviennent)*

☐ tout public intéressé, journalistes
☐ scientifiques ou étudiants diplômés d'autres disciplines
☒ les scientifiques de mon domaine (linguistique, pédagogie)
☒ autorités suisses, établissements

Le projet terminé, faut-il prévoir un embargo jusqu'à la mise en ligne des données ?

☐ oui ☒ non

Si oui, veuillez préciser :

Seront-ils nécessaires des logiciels rares ou coûteux pour une réutilisation des données ?

☒ oui ☐ non

Si oui, lesquels ? Il y a des lemmatiseurs qui coûtent quelque chose, mais je ne suis pas convaincue de les acheter car je n'en ai pas trouvé pour les trois langues

Est-ce que les données stockées dans l'archive du CSP peuvent être publiées chez FORS (http://forscenter.ch/) ou sur d'autres plateformes ?

☒ oui ☐ non

Détails : Nous aimerions publier notre corpus sur un site prévu pour des corpus linguistiques (ex. FORS, TMX ou Ortolang). Dans l'idéal, il pourrait s'agir d'une plateforme suisse ou spécialisée dans les corpus écrits d'apprenants

Organisation et traitement des données

Comment et où les données sont-elles stockées pendant la durée du projet ?
(Veuillez choisir toutes les réponses qui vous conviennent)

☒ serveur de l'Institut de plurilinguisme ☒ autres serveurs, services Cloud (ex. Dropbox)
☒ sous forme de papier (gardé à l'Institut) ☐ CD-R, disques durs externes ou supports similaires
☐ autre :

Détails : *Cliquez ici pour saisir du texte.*

Combien de fois et sous quelle forme une sauvegarde est-elle faite ? *(En dehors de la sauvegarde automatique du serveur de l'Institut de plurilinguisme)*

Jamais

Comment les données sont-elles décrites et documentées tout au long du projet ?

☒ selon le guide des bonnes pratiques du CSP pour la curation des données

☐ autre :

Quelles ressources (humaines, économiques, temps) sont mises à disposition pour l'organisation et la description des données ?

Collaborateur scientifique à 50% pendant 2 ans
Collaborateur scientifique à 30-40% pendant 5 mois (développement de l'instrument de calcul)
Aides-étudiants pour la lemmatisation/annotation ponctuelle (100-150 jours)

Divers

Est-ce qu'il y a d'autres informations sur l'organisation, le stockage et la documentation des données, qui pourraient concerner le centre de documentation du CPS ?

Cliquez ici pour saisir du texte.