C O R L I
Consortium Corpus, Langues et Interactions
TABLE RONDE
« Critères d'évaluation des corpus »

Evaluer la qualité des corpus ?
Pourquoi ? Comment ?

o.baude@paris10.fr

Remarques

- Un problème complexe…
- Une difficulté épistémologique
  - Place de la preuve dans la démarche scientifique
  - Place du corpus dans la démarche scientifique
  - Comparaison “Données” et “Publications”
  - Corpus = objets scientifiques très différents (de la linguistique descriptive à la linguistique appliquée)
  - En 2016 Corpus = corpus numérique = Facilité de constitution - Diffusion - Réutilisation - Transformation (Web de données)
  - Le web de données est (peut-être) un nouvel objet de savoir dont une partie repose sur la qualité des corpus

- Un cadre propice à la profusion de corpus (diffusion de la recherche)
  - Loi de 1982 sur la diffusion de la recherche
  - L’UE en 2001 prône l’interopérabilité des données (puis open access)
  - 2003 *Déclaration de Berlin sur le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales*,
  - 2016 Loi pour une république numérique (libre accès).
  - Depuis 2005, des plateformes, un Equipex, une TGIR

- Un constat = place croissante des corpus / objet varié / dont la qualité est en enjeu majeur

Pourquoi les évaluer ?

- **Comme production scientifique d'un chercheur**
  - Qualité de la preuve, de la démarche, de l'analyse
  - Marqueur (quantitatif et qualitatif) de l'activité scientifique
- **Comme production scientifique d'un laboratoire, d'une équipe, d'une communauté**
  - Marqueur de l'activité
  - Nécessite une quantification de la part de chaque acteur
- **Comme objet « patrimonial »**
  - Qui doit être archivé
  - Qu'on peut réutiliser
  - Qui contribue à une légitimité

Comment évaluer un corpus ?

- Un objet qu’on peut **quantifier**
- Un objet qu’on peut **qualifier**

Comment évaluer un corpus ?

- Un objet qu'on peut **quantifier**
  - Taille :
    - Nombre de mots,
    - Durée pour des enregistrements audio/video,
    - Type et quantité d'annotations
  - Temps passé / coût ?

Comment évaluer un corpus ?

- Un objet qu'on peut **qualifier**
  - Par l'expertise des pairs :
    - Corpus de thèse
    - Associé à un programme de recherche
    - Accepté sur une plateforme de gestion
  - ¨Par les usages au sein d'une discipline et hors de la discipline
    - Citation
    - Nombre d'utilisateurs
    - Réutilisation
    - Bref en étant **accessible**

Comment évaluer un corpus?

- Pour être un objet (semi) autonome reconnu comme production scientifique il doit être **accessible**.
  - Dans le cas contraire, c'est la production scientifique secondaire (thèse, article, livre, communication,…) qui seule sera évaluée.
  - *Proposition pour une position radicale : ne seront évalués que les corpus* ***accessibles***
    - Cela ne va pas de soi (exemples)
    - Choix scientifique et politique (open access)
    - Cas particulier des corpus nécessitant un restriction d'accès

Un corpus accessible ?

- **Conservé**
  - La dynamique et le stade du processus de conservation est un indicateur
    - Du stockage sécurisé à l'archivage
    - **Démarche de dépôt sur une plateforme reconnue (Cocoon, Ortolang, Huma-Num, Hal, autres ?)**

Un corpus accessible ?

- **Identifié**
  - Identifiant unique et pérenne
    - (attribué par les plateformes : Cocoon, Ortolang, Huma-Num (Nakala), centre Clarin, Hal,…)

- **Décrit** (cf. interventions sur la description)
  - Métadonnées
  - Interopérabilité des métadonnées
  - Finesse de description
    - Décrire le corpus et décrire les éléments de la collection (documents)
  - Guide de bonnes pratiques
  - Fonctionnalité au sein et par les plateformes
  - Qualité des métadonnées

Un corpus accessible ?

- **Aspects juridiques**

  - Protection de la vie privée et données personnelles
    - Consentement (CNIL, Bonnes pratiques)
    - Anonymisation (Bonnes pratiques, outils)
    - Repérage et traitement des restrictions d'accès (justifiées) (SIAF, plateformes, encore du flou dans les pratiques)

  - Respect du droit d'auteur
    - Privilégier les corpus pouvant être diffusés et réutilisés
    - Avoir prévu la gestion des droits le cas échéant (licences, chartes)
    - Paternité (références pour les citations)

Un corpus accessible ?

- Degré d'**interopérabilité**
  - Formats
  - Codages
  - Catalogage
  - Possibilité d'enrichissements

  - Guide de bonnes pratiques
  - Fonctionnalité au sein et par les plateformes

Un corpus accessible ?

- Un corpus **signalé**
  - Cf. points précédents
  - **Démarche de dépôt** sur une plateforme qui permet le signalement

Propositions

- 2 niveaux d'évaluation :
  - Accessibilité et interopérabilité
    - Pratiques partagées (Hal, MediHal, Clarin,…) : les plateformes se prononcent sur les aspects techniques et juridiques. Le dépôt dans une plateforme (et le degré d'appropriation des fonctionnalités de la plateforme) est un critère d'évaluation
  - Evaluation par les pairs
    - Pratiques partagées avec les revues (diffusion après évaluation par un comité, nombre de citation, de réutilisation, poids dans le champ…)

Propositions

- Besoins:
  - d'une communauté qui produit des bonnes pratiques, des formats, des outils, etc.
  - d'une communauté qui situe les enjeux scientifiques dans une démarche épistémologique,
  - d'une communauté qui dialogue avec d'autres communautés,
    - la qualité de certains aspects des traitements de corpus (description, indexation,…) dépassent les compétences du linguiste (métier de la documentation et de l'information scientifique),
    - Beaucoup de questions méthodologiques, technologiques et scientifiques traversent d'autres disciplines des SHS (et au-delà)
  - de plateformes « pérennes »