Pourquoi déposer ses corpus ?

Déposer un corpus c'est une participation au patrimoine

- Raison la plus souvent invoquée
  - Les corpus sont des données rares et uniques
    - Surtout vrai dans le cas de recueils originaux
    - La rareté peut dépendre de l'unicité des conditions de recueil mais aussi des coûts ou des personnes impliquées dans le recueil
  - L'évaluation peut porter sur ces critères
    - Impossibilité de reproduire un recueil
    - Coût du recueil

Déposer un corpus c'est un besoin scientifique

- Les sciences en général sont de plus en plus souvent accusées de fraude, de plagiat, ou simplement amènent à des interrogations, des remises en cause, …
  - Il faut donc assurer le contrôle des travaux
  - Leur reproductibilité
  - Besoins non spécifiques des sciences du langage
- Pour cela il faut disposer de toutes les données qui ont amené à des résultats
  - y compris dans les études n'utilisant pas de corpus
  - **les corpus sont un des cas d'études linguistiques** (avec les études expérimentales) **qui permettent de contrôler ou reproduire le travail scientifique**

Evaluation d'un dépôt "scientifique"

- Le dépôt pour reproductibilité et contrôle d'un travail scientifique peut s'inspirer de ce qui est fait dans les articles scientifiques ou dans les autres sciences
  - Il faut pouvoir justifier:
    - De l'origine des données
    - De la manière dont elles sont traitées
    - Du format dans lesquelles elles sont données
  - Comme un travail expérimental doit pouvoir être reproduit, un corpus doit pouvoir être ré-analysé
    - Fournir les formats et éventuellement les outils
  - Il doit aussi pouvoir mener à des analyses nouvelles
- Conséquences à attendre sur la manière de décrire les données dans les articles scientifiques

Enrichir les bases de données scientifiques

- L'avancée des recherches amènent à demander des données de plus en plus vastes, de plus en plus variées, de plus en plus exhaustives.
- Le recueil de données est extrêmement couteux
  - Il faut donc mutualiser cet effort
- La mutualisation présente aussi l'avantage d'amener la variété

Evaluer pour l'enrichissement des données

- Est-ce que le format des données est correctement décrit ?
- Quelle taille font ces données ?
  - Critère qui peut être subjectif et difficile à mesurer
- Sont-elles compatibles avec d'autres données qu'elles enrichissent ?
- S'agit-il de données novatrices (nouvelles) ?
  - Lien avec la pérennisation

Conditions pour le dépôt et l'évaluation

- Ces conditions d'évaluation (pérennisation, contrôle, enrichissement, réutilisation) ont des conséquences sur les conditions de dépôts
- Fournir à la communauté les outils et les conditions nécessaires
- La communauté s'évalue elle-même et donc ces conditions et ces outils doivent évoluer avec les avancées des travaux