Research Data Alliance & les données linguistiques

Helene N. Andreassen
UiT Université arctique de Norvège

Andrea L. Berez-Kroeker
University of Hawai'i at Manoa

Lauren Gawne
La Trobe University

Journées FLOraL-PFC: dialectologie et phonologie de corpus, 23-25 novembre 2017, Paris

La reproductibilité de la recherche

Evolution du terme *replicabilité*, la réutilisation d'une méthode scientifique afin de collecter de nouvelles données qui pourront confirmer des affirmations existantes.

Si la replicabilité est impossible, comment tester la crédibilité de la recherche?

L'accès aux données originales autorise une ré-analyse indépendante.

(Berez-Kroeker et al., à paraître)

La reproductibilité de la recherche

La reproductibilité nécessite

1. La transparence de la méthode appliquée lors de la collecte et l’analyse des données.

2. La transparence du statut des données, y compris le type de données, et si/comment le lecteur peut y avoir accès.

(Berez-Kroeker et al., à paraître)

Quid la réalité dans les publications?

Analyse de 50 grammaires et 50 thèses (2003-2012)

1. Méthode

\+ Période de collecte, type de données, participants

÷ Outils, équipement, logiciels, archivage

(Gawne et al., 2017, p. 172)

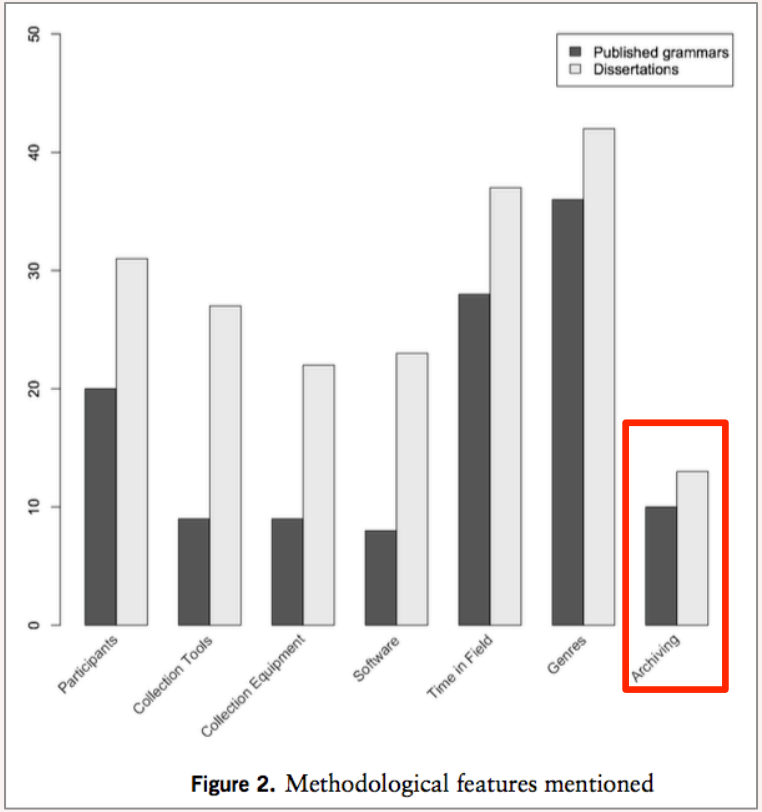


Figure 2. Methodological features mentioned

Quid la réalité dans les publications?

2. Conventions de citation, exemples numérotés

La majorité ne donne pas de citation à côté des exemples.

Un certain nombre de thèses réfère au corpus, archivé ou pas.

(Gawne et al., 2017, p. 175)

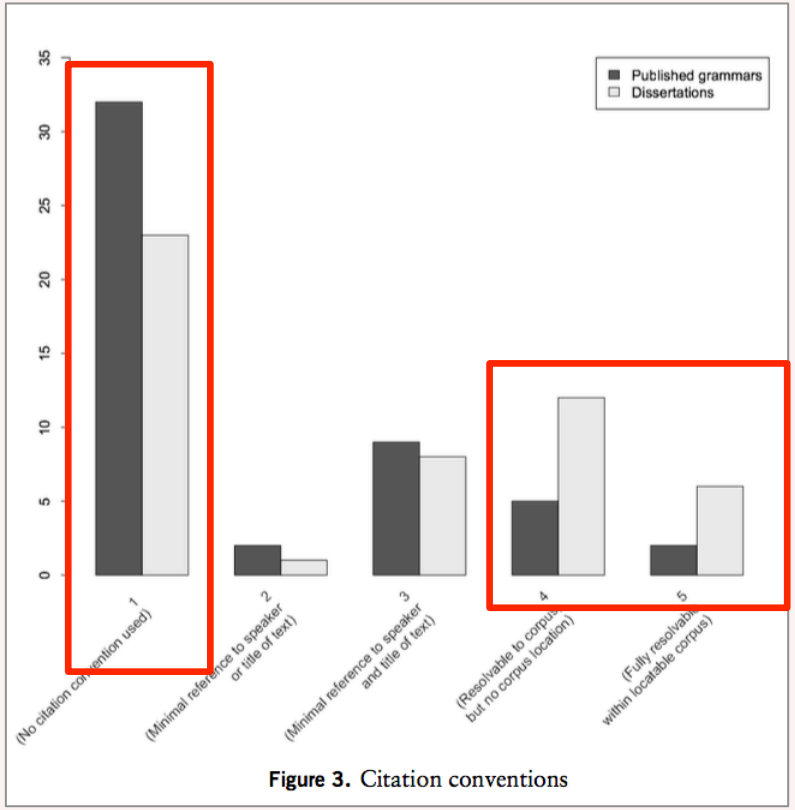


Figure 3. Citation conventions

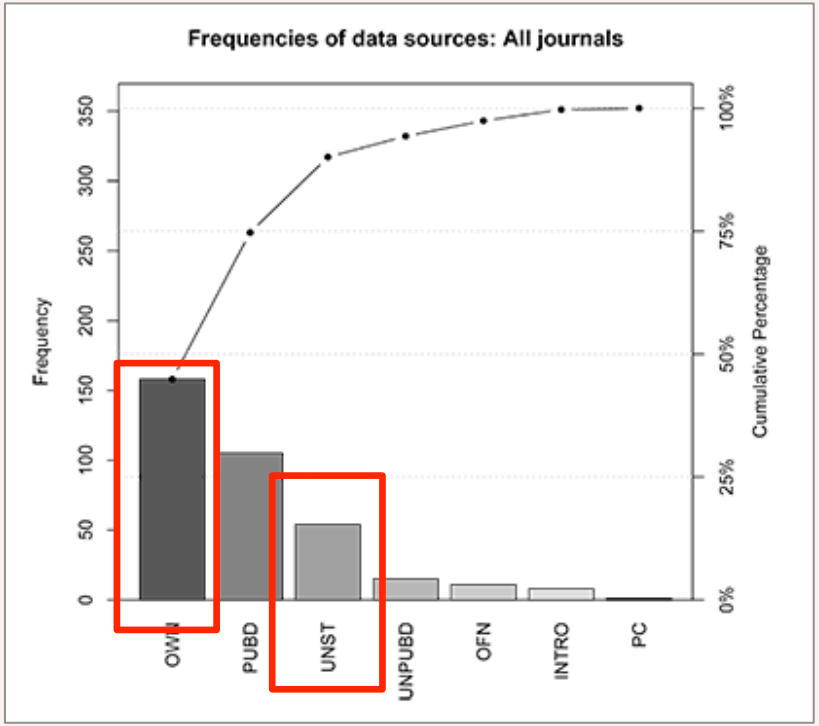
Quid la réalité dans les publications?

Analyse de 9 revues en linguistique (2003-2012)

1. Source

Préférence pour l'usage de données collectées par l'auteur lui-même.

Un certain nombre de publications ne mentionne pas la source des données.



(Berez-Kroeker et al., 2017b)

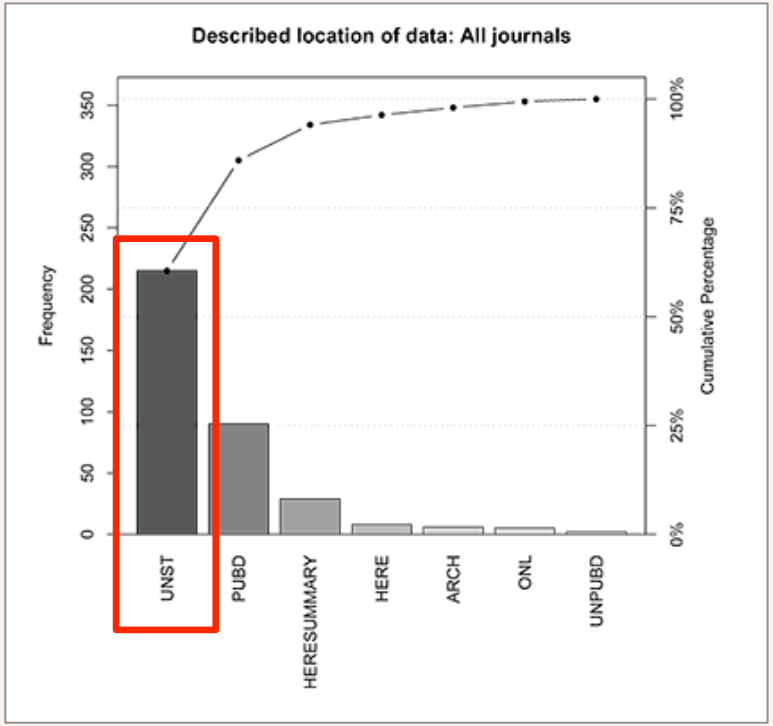
Quid la réalité dans les publications?

2. Emplacement

Plus de la moitié n'explicite pas l'emplacement des données.

Egalement absence de mention de collecte à travers introspection.

Observation positive: Les différentes sous-disciplines favorisent la transparence pour des éléments concrets.



(Berez-Kroeker et al., 2017b)

Une des initiatives

Projet NSF *Developing Standards for Data Citation and Attribution in Linguistics*: plus de 40 participants internationaux, 3 workshops, session de panel lors de la rencontre annuelle de LSA en 2017.

Résultats:

- 2 enquêtes grammaires et revues (Berez-Kroeker et al., 2017b, Gawne et al., 2017)
- Papier de position sur la citation de données linguistiques (Berez-Kroeker et al., à paraître)
- 1ère version de principes de citation de données linguistiques
- Création de LDIG

Linguistics Data Interest Group (LDIG)

Objectifs:

- Développement et adoption de principes et guidelines pour la citation et l'attribution des données (chercheurs, organisations, éditeurs, archives)
- Education et dissémination (formation et sensibilisation, changement de culture)
- Valorisation du traitement des données dans la profession (traiter le travail concernant des données comme un output scientifique)

https://www.rd-alliance.org/groups/linguistics-data-ig

Research Data Alliance (RDA)

Organisation qui vise à créer une infrastructure sociale et technique afin de faciliter le partage libre de données scientifiques.

Plus de 6000 membres venant de 132 pays, rencontres 2 fois par an.

Groupes thématiques:

- *Interest groups*: Plateforme internationale de communication et coordination pour des individus avec les mêmes intérêts, à l'intérieur et à l'extérieur de RDA; production de recommandations et rapports; création de Working Groups. Ex: Metadata IG, PID IG.

- *Working Groups*: Réseau international dont le but est de faire avancer le travail de manière très concrète. Ex: Data Citation WG, Data Versioning WG.

The Austin Principles of Data Citation in Linguistics

Objectifs:

- Encourager et améliorer la visibilité et la récupérabilité des données
- Guidelines pour le formatage de citations de données
- Etre applicable à toute sous-discipline ainsi que tout type de données
- Format:
    - Standard, up-to-date (best practice)
    - Lisibilité machine et humaine

Modèle:

- Interprétation des principes de FORCE11 (2014)

(Berez-Kroeker et al., 2017a)

The Austin Principles of Data Citation in Linguistics

1. Importance

   Les données devraient être traitées comme des produits de recherche légitimes et citables.

   En linguistique, les données peuvent également être un documentation d'héritage culturel, d'évolution sociétale et de potentiel humain.

(Berez-Kroeker et al., 2017a, traduction: Andreassen)

The Austin Principles of Data Citation in Linguistics

2. Crédit et attribution

   Les citations des données devraient faciliter l'accord de crédit scientifique et d'attribution normative et légale.

   En linguistique, cela peut être appliqué à tout individu qui participe dans la collecte/création des données (p.ex. les locuteurs natifs, les interviewés, les transcripteurs).

(Berez-Kroeker et al., 2017a, traduction: Andreassen)

The Austin Principles of Data Citation in Linguistics

3. Preuve

   Dans la littérature scientifique, lorsqu'une affirmation dépend de données, celles-ci devraient être citées.

   En linguistique, les données devraient être accessibles, et la méthode de collecte et d'analyse de données devrait être transparente.

(Berez-Kroeker et al., 2017a, traduction: Andreassen)

The Austin Principles of Data Citation in Linguistics

4. Identification unique

   Une citation de données devrait inclure un type d'identification pérenne, unique et largement utilisée.

   Dans la sélection d'archive, le linguiste devrait chercher une archive qui offre une identification sous la forme d'un identifiant pérenne (PID).

(Berez-Kroeker et al., 2017a, traduction: Andreassen)

The Austin Principles of Data Citation in Linguistics

5. Accès

   Les citations de données devraient faciliter l'accès, lisible par les machines et les êtres humains, aux données et à des informations apparentées.

   Les données devraient être aussi accessibles que possible, afin d'assure la reproductibilité, et si fermées que nécessaire, pour respecter les contraintes éthiques, légales et communautaires.

(Berez-Kroeker et al., 2017a, traduction: Andreassen)

The Austin Principles of Data Citation in Linguistics

6. Pérennité

   Les PIDs, métadonnées et les données devraient persister.

   Les linguistes devraient stocker les données dans des archives ayant une politique qui déclare la pérennité des données, métadonnées et identifiants.

(Berez-Kroeker et al., 2017a, traduction: Andreassen)

The Austin Principles of Data Citation in Linguistics

7. Spécificité et vérifiabilité

   Les citations de données devraient faciliter l'identification et la vérifiabilité des données particulières derrière une affirmation, ainsi que l'accès à celles-ci.

   La citation devrait faciliter la récupérabilité d'un point précis dans les données qui est derrière une affirmation, avec une méthode systématique d'identification.
   La citation devrait refléter la nature non-statique des données et inclure la version.

(Berez-Kroeker et al., 2017a, traduction: Andreassen)

The Austin Principles of Data Citation in Linguistics

8. Interopérabilité et flexibilité

   Les méthodes de citation de données devraient être suffisamment flexibles, mais pas plus que nécessaire, afin d'accommoder la variation de pratiques dans les communautés.

   Les linguistes travaillent sur de différents types de données, et les standards de citation devraient refléter ce fait, sans aller contre ces principes.
   Nous encourageons les éditeurs de développer de formats de citations ainsi qu'une politique sur la base de ce document.

(Berez-Kroeker et al., 2017a, traduction: Andreassen)

Linguistics Data Citation

linguisticsdatacitation.org

Linguistics Data Citation

Endorse

Data citation will only become standard practice in linguistics if, as individual researchers, we commit to improving data citation in our own work, and in our specific research communities.

By endorsing the Austin Principles you show other linguists that you support reproducible research. You can join the list of linguists who endorse the principles by adding your name to this Google Form.

If you would like to be more actively involved in shaping the future of linguistic data, we encourage you to join the Linguistics Data Interest Group as part of the Research Data Alliance.

***

We, the undersigned, endorse the Austin Principles of Data Citation in Linguistics. We support the idea that the data on which linguistic analyses are based are of fundamental importance to the field, and should be treated as such.

List of Endorsers (last updated November 20, 2017)

Prochaine étape

Les citations en linguistique et les métadonnées standards

Workshop pour commencer la discussion:

RDA 11th Plenary Meeting
Berlin, mars-18

Bibliographie

Berez-Kroeker, A. L., Andreassen, H. N., Gawne, L., Holton, G., Kung, S. S., Pulsifer, P., Collister, L. B., The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. (2017a). *Draft: The Austin Principles of Data Citation in Linguistics (Version 0.1).* http://site.uit.no/linguisticsdatacitation/austinprinciples/ Accédé 19.11.2017.

Berez-Kroeker, A. L., Gawne, L., Kelly, B. F., & Heston, T. (2017b). *A survey of current reproducibility practices in linguistics journals, 2003-2012.* *https://sites.google.com/a/hawaii.edu/data-citation/survey*

Berez-Kroeker, A. L., Gawne, L, Kung, S. S., Kelly, B. F., Heston, T., . . .Beaver, D. I. (à paraître). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics.*

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles.* Martone M. (ed.) San Diego CA: FORCE11. https://doi.org/10.25490/a97f-egyk

Gawne, L., Kelly, B. F., Berez-Kroeker, A. L., & Heston, T. (2017). Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation, 11*, 157-189.

Research Data Alliance & les données linguistiques

Helene N. Andreassen, Andrea L. Berez-Kroeker & Lauren Gawne

RESEARCH DATA ALLIANCE

AUSTRALIAN NATIONAL DATA SERVICE

UiT / THE ARCTIC UNIVERSITY OF NORWAY