Données IPFC : base et référencement bibliographique

Isabelle Racine, Sylvain Detey & Helene N. Andreassen

ELCF, U. de Genève & U. Waseda & UiT Université Arctique de Norvège

Journées Floral-(I)PFC
Paris, MSH, 26-27 novembre 2018

WASEDA University

UiT THE ARCTIC UNIVERSITY OF NORWAY

UNIVERSITÉ DE GENÈVE

FNSNF FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE

Pour l'instant…

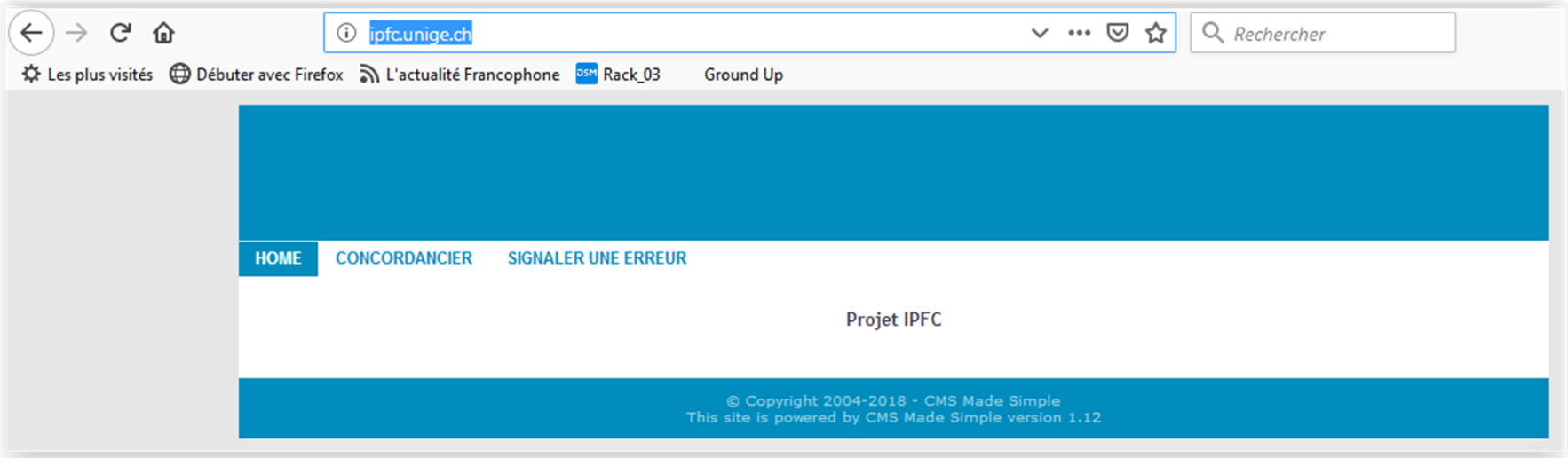
Un site (http://cblle.tufs.ac.jp/ipfc/)

Objectifs :

- **Présenter le projet**
- **Présenter les différentes équipes du projet (1 page par équipe)**
- **Présenter les colloques liés au projet et les actualités**

Pour l'instant…

Une base de données en phase pilote (**http://ipfc.unige.ch/**)



- Créée et développée par Pierre Ménétrey grâce à deux financements de l'Université de Genève (2015 et 2018).

- Voir présentation dans les Journées IPFC de 2016.

Pourquoi une base de données IPFC?

- Pour que les données IPFC acquièrent le statut de corpus phonologique

> "… a collection of recordings which are available in a **computer-readable form** (e.g. wav format) and which are **accompanied by transcriptions and annotations aligned with the signal**. The transcriptions and annotations should be in standardized formats […] or in formats easily convertible to them […]. They should contain **essential metadata**: information about how and when the recordings were made, how the speakers were selected and who the speakers are (age, sex, social status, etc.). The transcriptions and annotations should be accompanied by a documentation explaining how they were devised. **All these requirements should be met if a corpus is to be searchable** so that analyses made by other users of the corpus can be verified and (in)validated. Finally, the collection of the data and its availability for users must follow **agreed ethical guidelines** which can vary from country to country."
>
> Detey, Durand, Laks & Lyche (2016)

- Voir également Gut (2014).

Pourquoi une base de données IPFC?

- Deux objectifs avaient été mentionnés en 2016:
  - Assurer une pérennisation des données IPFC collectées
  - Faciliter les recherches dans les données

    ⇨ par le biais d'un concordancier permettant des recherches par mot-clé en fonction de la L1, du locuteur ou du type de tâche (+ critères liés au locuteur: p. ex. durée des études).

- Mais un troisième objectif s'ajoute:
  - Nécessité de pouvoir faire référence aux **données** elles-mêmes et pas seulement aux publications!

Pourquoi citer les données IPFC?

- La citation des données dont on se sert permet au lecteur de les consulter lui-même.
- Elle permet également de présenter l'auteur des données et les conditions de réutilisation.
- Elle permet de faire le lien entre le papier et les données.

Ressources:

Ball, A., & Duke, M. (2015). How to cite datasets and link to publications. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides

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. (2018). *The Austin Principles of Data Citation in Linguistics*. Version 1.0. http://site.uit.no/linguisticsdatacitation/austinprinciples/

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., . . . Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics, 56*(1), 1-18. https://doi.org/10.1515/ling-2017-0032

La référence bibliographique

Élément obligatoire	Explication	Exemple
Auteur	Le créateur de l'ensemble de données (= dataset).	Meisenburg, T.
Date de publication	Le moment où les données, ou les métadonnées, sont mises en accès libre. Si période d'embargo sur les fichiers, mettre le moment où l'embargo expire.	2002
Titre	Nom de l'ensemble de données.	Enquête Lacaune
Emplacement	Identifiant persistant, p.ex. doi. Si pas d'identifiant persistant, mettre l'url de la page principale de la collection.	https://public.projet-pfc.net/
Éditeur	Nom de l'archive ou l'organisation qui accueille (et qui assure la qualité de) les données.	Base PFC publique

Élément optionnel	Explication	Exemple
Version	Si les données ont été modifiées, le numéro de version change. Si pas de version indiquée, mettre la date de téléchargement.	01.11.2018
Identifiant apparenté	Si l'ensemble de données fait partie d'une collection plus grande, mettre le nom de la collection.	Phonologie du français contemporain

Référence bibliographique:

Exemple de (méta) données non publiées

Andreassen, H. N. & Lyche, C. (non publié). Enquête Tromsø. *Interphonologie du français contemporain*. http://cblle.tufs.ac.jp/ipfc/

A terme:

Métadonnées publiées mais données pas en accès libre:

Andreassen, H. N. & Lyche, C. (2019). Enquête Tromsø. *Base IPFC*. http://ipfc.unige.ch/

Etape finale (ou pas?): données publiées en accès libre:

Andreassen, H. N. & Lyche, C. (2019). Enquête Tromsø. *Base IPFC*. Téléchargé le ? ??? 20?? de http://ipfc.unige.ch/

+ voir si ajouter une indication de la version des données.

Citations dans le texte – exemples:

1. Maintenant je cite l'ensemble de données (Meisenburg, 2002).
2. Maintenant je cite un fichier particulier dans l'ensemble de données (Meisenburg, 2002, nom de fichier: 81aaag_anon_wav).
3. Maintenant je donne un exemple numéroté:

   1) *Je me suis retrouvé avec des gens qui venaient de la Martinique.* (81aaag_anon_wav, 1:40)

   NB! Pour les exemples numérotés, si pas évident à partir du contexte, mettre le nom d'auteur et la date de publication devant le nom du fichier.

Les difficultés

- Financières et techniques:
  - Difficile de s'assurer la contribution d'un ingénieur informatique. De manière permanence = impossible et de manière suffisamment longue pour permettre un développement qui ne relève pas du casse-tête = très difficile

- L'intégration des données dans la base:
  - Qui peut faire la saisie (accès au serveur de l'Université de Genève en externe)?
  - Données minimales nécessaire: métadonnées (locuteur + enquête) + 2 fichiers par tâche et par locuteur (son + grid avec transcription orthographique selon conventions IPFC mais SANS CODAGES)
  - Seules des données vérifiées et anonymisées (en fonction des exigences de chaque pays) seront entrées dans la base
  - Un numéro DOI par enquête (avec renvoi aux références à citer pour l'utilisation de chaque point d'enquête) + le choix d'une licence de type Creative Commons

- Les accès:

Tension entre la tendance à l'Open Access (exigée par certains organismes de financement!) et la Protection personnelle des données (exigée par certains comités d'éthique qui sont à convaincre avant le début d'une enquête!)

Les accès proposés en 2016 et en développement

1. A court terme, un accès public à des données très restreintes:

   - Un accès via le concordancier – donc uniquement par le biais d'une recherche lexicale – aux extraits de productions D'UN SEUL apprenant par enquête et à des métadonnées restreintes (sexe, âge, L1, études de français + infos sur l'enquête)
   - Possibilité de télécharger l'extrait (fichier son et grid de l'extrait)

   - Mais attention, nécessité d'indiquer suffisamment clairement dans le concordancier comment faire référence à ces données!

   - Permet de rendre visible le projet et à des personnes externes d'utiliser un extrait à des fins didactiques, ce qui nécessite l'association d'une licence de type Creative Commons (mais réfléchir à laquelle).

Les accès proposés en 2016 et en développement

2. A court terme, un accès «recherche» aux membres de chaque équipe:

- Un accès via le concordancier – donc uniquement par le biais d'une recherche lexicale – aux extraits de productions de tous les apprenants d'une ou plusieurs enquêtes et aux métadonnées.
- Possibilité de télécharger les extraits (fichiers son et grid des extraits) mais pas les fichiers sons complets!

- Qui pourrait bénéficier d'un tel accès?
  - Un étudiant/doctorant qui travaille sur des données qu'il n'a pas (entièrement) collectées et qui est supervisé par un responsable/membre d'une équipe
  - Les membres de chaque équipe

Proposition:

- Demander un accès via un formulaire qui permet de cocher chaque enquête pour laquelle un accès est souhaité.
- Demande via un mail générique adressée à Genève puis Genève s'assure de l'accord du responsable d'équipe/d'enquête avant de générer un login personnalisé
- Accès via un login personnalisé dont la durée est limitée dans le temps

Les accès proposés en 2016 et en développement

3. A court terme, un accès «admin» aux responsables d'équipe/enquête:
   - Un accès permettant d'entrer et de gérer les données (métadonnées et fichiers son et grid)

4. A plus long terme et pour les enquêtes pour lesquelles la convention signée le permet, un accès public complet (concordancier + téléchargement des données et des métadonnées complètes) à une enquête

Exemple – le projet PFC – base publique

Base PFC publique

PFC public database

17 enquêtes anonymisées en ligne

[Enquêtes] [Transcriptions] [Liaisons] [Schwas]

- Abidjan (cia) Info
- Aix-Marseille (13b) Info
- Brécey (50a) Info
- Burkina Faso (bfa) Info
- Dijon (21a) Info
- Domfrontais (61a) Info
- Lacaune (81a) Info
- Nantes (44a) Info
- Neuchâtel (sca) Info

Enquête Neuchâtel

- **Pays :** Suisse
- **Région :** Neuchâtel
- **Responsable :**
- **Travail de terrain :** Nathalie Bühler, Isabelle Racine
- **Transcriptions :** Isabelle Racine, Françoise Zay, Jean-Paul Philippe et Nathalie Bühler
- **Codages :** Isabelle Racine, Helene N. Andreassen
- **Vérification :** Isabelle Racine, Helene N. Andreassen
- **Date :** 2011-09-10
- **Publication principale :** Racine Isabelle/Andreassen, Helene (2012): "A phonological study of a Swiss French variety : data from the Canton of Neuchâtel", in: Gess, Randall/Lyche, Chantal/Meisenburg, Trudel (eds.): Phonological variation in French : Illustrations from three continents, Amsterdam : John Benjamins, 173-207.
- **Autres publications :** Andreassen, H., Maître, R., Racine, I. (2010). La Suisse. Dans S. Detey, J. Durand, B. Laks et C. Lyche (éds). Les variétés du français parlé dans l'espace francophone : ressources pour l'enseignement. Paris : Ophrys, 211-233. Andreassen, H & Racine, I. (2016). Variation in Switzerland: the behaviour of schwa in Martigny, Neuchâtel and Nyon. In S. Detey, J. Durand, B. Laks et C. Lyche (eds), Varieties of Spoken French. Oxford: Oxford University Press, 430-440. Avanzi, M., Schwab, S. & Racine, I. (2015). A Preliminary Study of Penultimate Accentuation in French. In J. Romero & M. Riera (eds). The Phonetics-Phonology Interface. Representations and methodologies, Amsterdam : John Benjamins: 93-107. Racine, I. (2016). Le français en Suisse. In : S. Detey, I. Racine, Y. Kawaguchi & J. Eychenne (eds). La prononciation du français dans le monde : du natif à l'apprenant. Paris : CLE International (coll. Didactique des langues étrangères), 44-48. Racine, I., Andreassen, H & Benetti, L. (2016). Swiss French. In S. Detey, J. Durand, B. Laks et C. Lyche (eds), Varieties of Spoken French. Oxford : Oxford University Press, 223-235. Racine, I., Durand, J. & Andreassen, H. N. (2016). PFC, codages et représentations : la question du schwa, Corpus, 15, Corpus de français parlé et français parlé des corpus, 213-236. Racine, I., Schwab, S. & Detey, S. (2013). Accent(s) suisse(s) ou standard(s) suisse(s) ? Approche perceptive dans quatre régions de Suisse romande. Dans A. Falkert (éd.). La perception des accents du français hors de France. Mons : Editions CIPA, 41-59. Schwab, S. & Racine, I. (2012). Le débit lent des Suisses romands : mythe ou réalité? Journal of French Language Studies, 23 (2), 281-295.
- Description de l'enquête

- **Locuteurs:**
  - scaaf1 Sexe : M | Age : 78 | Info | Transcriptions | Schwas | Liaisons | Telecharger les données sur Ortolang | Ecoute synchronisée
  - scajc1 Sexe : F | Age : 27 | Info | Transcriptions | Schwas | Liaisons | Telecharger les données sur Ortolang | Ecoute synchronisée
  - scajb2 Sexe : M | Age : 31 | Info | Transcriptions | Schwas | Liaisons | Telecharger les données sur Ortolang | Ecoute synchronisée
  - scajb1 Sexe : F | Age : 78 | Info | Transcriptions | Schwas | Liaisons | Telecharger les données sur Ortolang | Ecoute synchronisée
  - scahd1 Sexe : F | Age : 54 | Info | Transcriptions | Schwas | Liaisons | Telecharger les données sur Ortolang | Ecoute synchronisée
  - scacy1 Sexe : F | Age : 43 | Info | Transcriptions | Schwas | Liaisons | Telecharger les données sur Ortolang | Ecoute synchronisée

Toute utilisation des données PFC d'un ou deux points d'enquête particuliers doit être accompagnée d'une référence à l'enquête ainsi qu'à sa publication principale (indiquée dans les métadonnées de l'enquête)

Conclusion

- Un GROS chantier….
- Pour lequel un financement est difficile à trouver
- Qui pose des questions fondamentales
- Qui cristallise cette tension entre Open Access et PDP

Merci de votre attention!