
LNG-1100 : Méthodes expérimentales et analyse de données

Question de recherche, collecte de données, intro à R

Guilherme D. Garcia

fr.gdgarcia.ca

2



Plan de la séance

1. Question de recherche et collecte de données
2. Intro à R (Barnier, 2023, chapitres 1–3; 5–7)
3. Pratique



Question de recherche

☞ Vos questions déterminent vos méthodes (collecte + analyse)

1. Choisissez un sujet qui vous intéresse
2. Développez des questions de recherche
 - **Examinez les études déjà publiées** (p. ex., vers la fin des articles)
 - Discutez avec vos collègues et profs
 - Explorez des données existantes et observez les patrons qui émergent

☞ Votre question ne doit pas être trop spécifique ni trop générale :



3. Considérez des résultats possibles et les étapes suivantes



Question de recherche

Exemples

Ordonnez les questions (spécifique → générale)

- A. Comment les locuteurs du français montréalais produisent-ils les voyelles nasales [ã] et [ɛ]^a en langage familier?
- B. Quelle est la variation exacte de la fréquence fondamentale (F_0)^b pour la voyelle [i] dans le mot “si” lorsqu’elle est prononcée par des locuteurs bilingues français-anglais montréalais?
- C. Qu'est-ce qui influence la langue?
- D. Quel est le rôle des processus phonologiques dans l'harmonie vocalique^c du finnois?
- E. Comment les facteurs sociaux affectent-ils le changement linguistique?

^aPar exemple, « temps » et « pain ».

^bLa fréquence initiale d'un son brut généré dans les cordes vocales.

^cLe processus dans lequel une voyelle devient plus similar à une autre voyelle : lotu → lutu.



Question de recherche

Exemples

Ordonnez les questions (spécifique → générale)

- B. Quelle est la variation de la fréquence fondamentale (F0) pour la voyelle [i] dans le mot « si » lorsqu'elle est prononcée par des locuteurs bilingues français-anglais montréalais?
- A. Comment les locuteurs du français montréalais produisent-ils les voyelles nasales [ã] et [ɛ] en langage familier?
- D. Quel est le rôle des processus phonologiques dans l'harmonie vocalique du finnois?
- E. Comment les facteurs sociaux affectent-ils le changement linguistique?
- C. Qu'est-ce qui influence la langue?



Collecte de données

Logiciels

- Plusieurs méthodes d'élicitation de données exigent des logiciels spécifiques
- Quelques options :

- [PsychoPy[↗]](#), [OpenSesame[↗]](#), [Praat[↗]](#)
- [Experiment builder[↗]](#), [E-prime[↗]](#), [MatLab[↗]](#), [SuperLab[↗]](#)

gratuits
\$\$\$

- En plus, **nombreuses** options pour des expériences en ligne

☞ Tout cela dépend de votre **question de recherche**



Bases de données

De quel type de données avez-vous besoin...?

- Écrites?
- Orales?
- Jugement?
- etc.



Bases de données

IRIS

- Base de données générale (plusieurs domaines/sujets) : iris-database.org ↗



Bases de données

CHILDES

Spécifique pour l'acquisition du langage¹ : childestalkbank.org ↗



CHILDES is the child language component of the [TalkBank](#) system.
TalkBank is a system for sharing and studying conversational interactions.

System	Database	Manuals
Ground Rules Contributing New Data IRB Principles Overviews and Introductions	**Index to Corpora** Browsable Database LuCiD Toolkit childe-db Hints on Downloading	CHAT - CLAN - MOR Tutorial Screencasts SLP's Guide to CLAN and 中文
Links	Programs	Contact
TalkBank Other Child Language sites Research based on CHILDES Child Language Diaries	CLAN - Example Files XML creator and XML Schema Related Software	Brian MacWhinney : homepage How to subscribe to Mailing Lists
Phonology and Fonts	Ideas	Morphology and Lexicon
Phon and PhonBank Unicode and IPA for Mac Unicode and IPA for Windows	Topics in language acquisition Teaching Resources Bibliographies Building a New Corpus	Part of Speech Analysis by MOR MRC lexical dictionary ChildFREQ Site and Paper

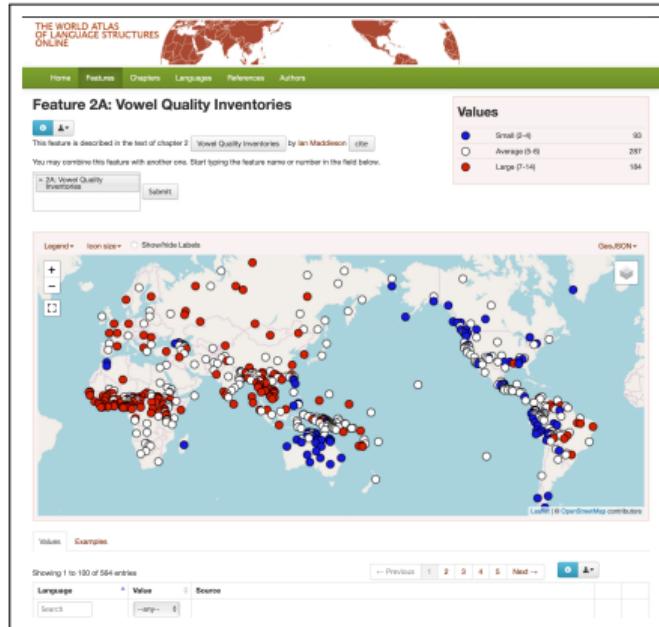
¹Une partie du projet [TalkBank](#) ↗.



Bases de données

WALS

Excellente base de données sur des traits linguistiques : wals.info↗



Bases de données

Speech accent archive

Les accents natifs ou non natifs de l'anglais : accent.gmu.edu ↗

The screenshot shows the homepage of the speech accent archive. At the top left is a stylized illustration of a human ear labeled "ear". To its right is a stylized illustration of lips labeled "mouth". In the center, the text "the speech *accent* archive" is displayed in a serif font. To the left of the main title is a vertical sidebar with the following navigation links: "how to", "browse", "search", "resources", and "about". Below these links are two more stylized illustrations: a brain labeled "brain" and a microphone labeled "microphone". To the right of the sidebar is a descriptive paragraph about the archive's purpose. At the bottom of the page, there is a link to practice phonetic transcription, the date of the last update (15 January 2019), and the number of samples (2780). Social media sharing icons for Facebook and Twitter are at the bottom right, along with the George Mason University logo.

the speech *accent* archive

how to
browse
search
resources
about

The speech accent archive uniformly presents a large set of speech samples from a variety of language backgrounds. Native and non-native speakers of English read the same paragraph and are carefully transcribed. The archive is used by people who wish to compare and analyze the accents of different English speakers.

practice your phonetic transcription for the speech accent archive: click here

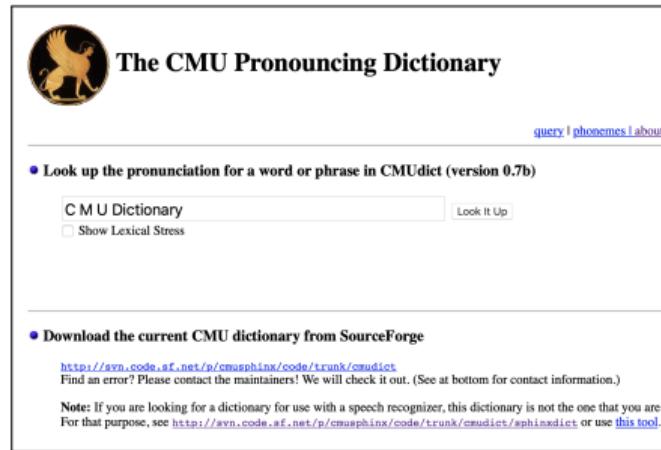
last updated: 15 january 2019 2780 samples



Bases de données

CMU Pronouncing Dictionary

- speech.cs.cmu.edu/cgi-bin/cmudict ↗



The screenshot shows the homepage of the CMU Pronouncing Dictionary. At the top is a logo of a golden griffin. Below it, the title "The CMU Pronouncing Dictionary" is displayed. A navigation bar at the top right includes links for "query", "phonemes", and "about". The main content area has two sections:

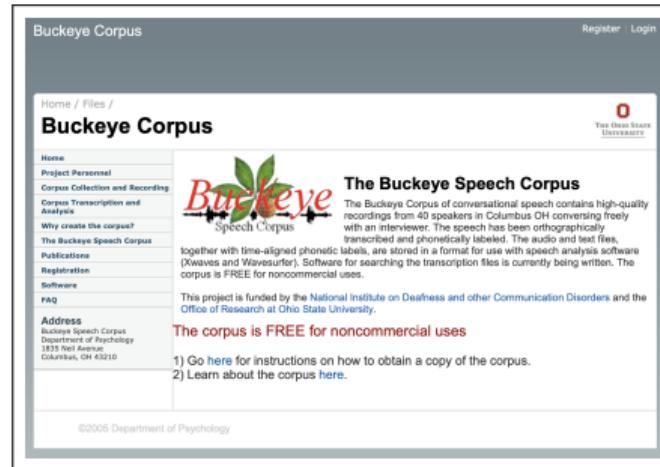
- Look up the pronunciation for a word or phrase in CMUDict (version 0.7b)**
A search input field containing "C M U Dictionary" with a "Look It Up" button next to it. There is also a checkbox labeled "Show Lexical Stress".
- Download the current CMU dictionary from SourceForge**
A link to the SourceForge repository: <http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict>.
Text below the link: "Find an error? Please contact the maintainers! We will check it out. (See at bottom for contact information.)"
A note at the bottom: "Note: If you are looking for a dictionary for use with a speech recognizer, this dictionary is not the one that you are For that purpose, see <http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/sphinxdict> or use [this tool](#)."



Bases de données

Buckeye corpus

Parole naturelle (avec transcription) du centre-ouest américain : buckeyecorpus.osu.edu ↗



The screenshot shows the homepage of the Buckeye Corpus. At the top, there's a navigation bar with links for "Home", "Files", "Buckeye Corpus", "Register", and "Login". On the right side of the header is a logo for "The Ohio State University" featuring a red 'O' and the university's name. The main content area has a large green "Buckeye" logo with a speech bubble icon. To the right of the logo, the text "The Buckeye Speech Corpus" is displayed. Below this, a detailed description of the corpus is given: "The Buckeye Corpus of conversational speech contains high-quality recordings from 40 speakers in Columbus OH conversing freely with an interviewer. The speech has been orthographically transcribed and phonetically labeled. The audio and text files, together with time-aligned phonetic labels, are stored in a format for use with speech analysis software (Xwaves and Wavewerker). Software for searching the transcription files is currently being written. The corpus is FREE for noncommercial uses." Further down, it states "This project is funded by the National Institute on Deafness and other Communication Disorders and the Office of Research at Ohio State University." Below that, a section titled "The corpus is FREE for noncommercial uses" lists two items: "1) Go [here](#) for instructions on how to obtain a copy of the corpus." and "2) Learn about the corpus [here](#)." At the bottom of the page, a copyright notice reads "©2005 Department of Psychology".

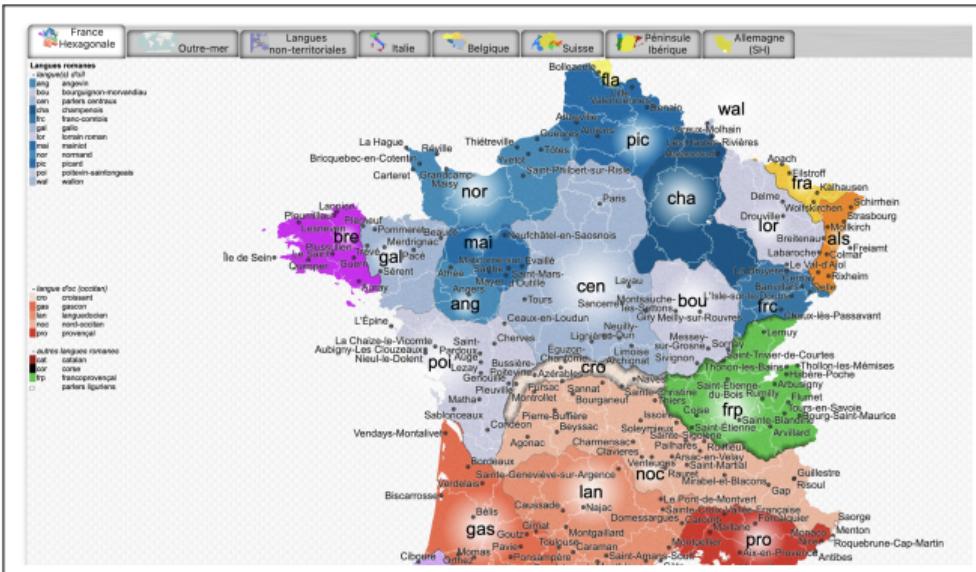
(Script pour échantillonner des mots à partir du corpus Buckeye)



Bases de données

Atlas sonore

- atlas.limsi.fr ↗



Bases de données

Corpus de français parlé au Québec

- applis.flsh.usherbrooke.ca/cfpq/

UNIVERSITÉ DE SHERBROOKE Centre d'analyse et de traitement informatique du français québécois Corpus de français parlé au Québec

Faculté des lettres et sciences humaines

CFPQ

Accueil Présentation Conventions Vue d'ensemble Renseignements Recherche mercredi 13 septembre 2023

ÉQUIPE Responsable ASSISTANTS Support technique Enregistrement Transcription et révision

Perdre ... ?
Ou ne pas perdre ce qu'on dit ?

Corpus de français parlé au Québec CFPQ

Corpus multimodal
Corpus qui intègre les trois dimensions caractéristiques d'une interaction verbale en face-à-face, à savoir ses dimensions verbale, paraverbale et gestuelle

STATISTIQUES
712,300 mots
28,638 mots différents

CONNEXION

Centre de recherche sur la société et la culture Québec

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Tous droits réservés © Université de Sherbrooke 2500, boul. de l'Université, Sherbrooke (Québec) CANADA J1K 2R1
Mise à jour le 23 janvier 2019 - Application développée avec cadrelic Yili(version 1.1.22)



15 sur 26

Intro à R



Questionnaire

forms.office.com/r/XgXnS2Y8wD



Pratique

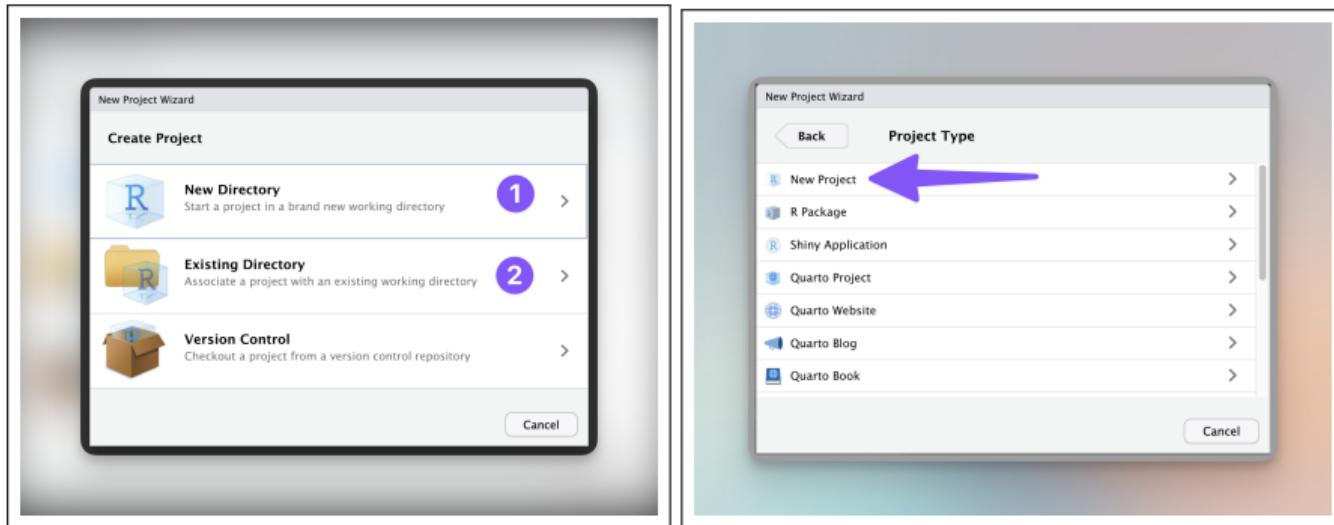
1. Créer un **projet R** pour notre cours
2. Organiser nos dossiers : consultez le tutoriel [ici](#)
3. Créer notre premier script pour aujourd’hui



Projet R pour LNG-1100

1. File > New Project...

Choisissez l'option 1 si vous n'avez pas encore un dossier pour le cours



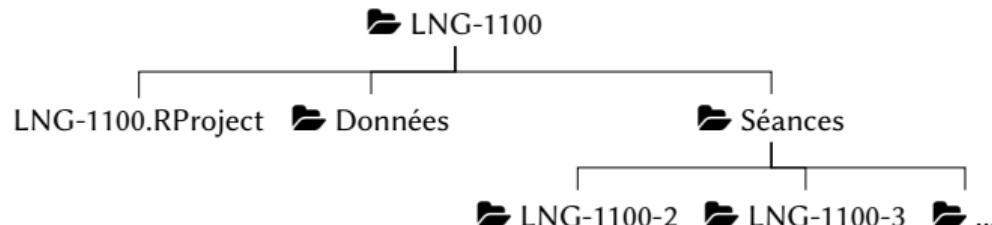
Projet R pour LNG-1100

Pourquoi un projet...?

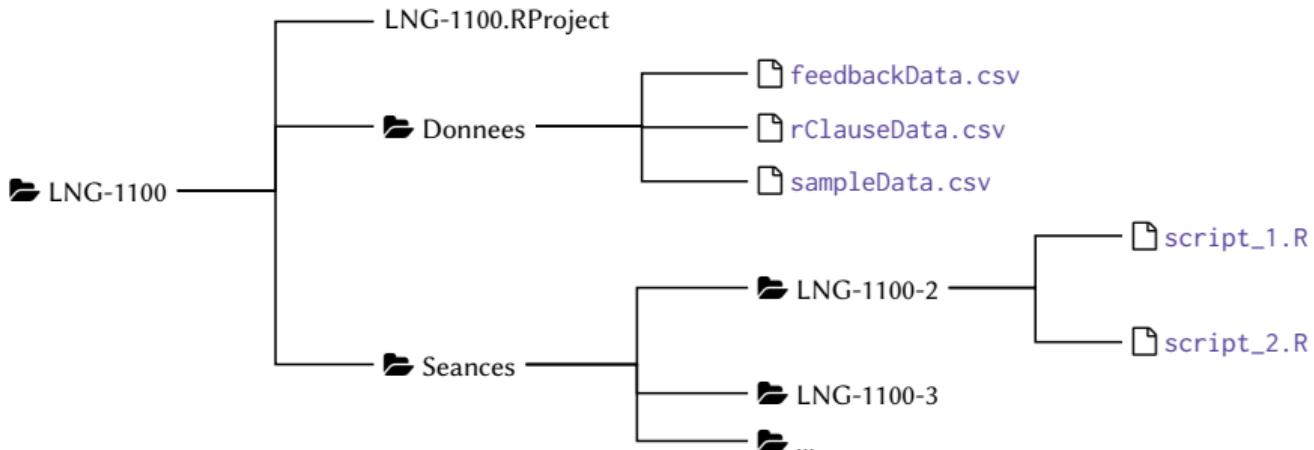
- On concentre tous les fichiers du cours dans un seul dossier
- RStudio connaîtra déjà la localisation des fichiers

Maintenant, on continue sur RStudio (consultez les fichiers plus tard)

☞ Voici une suggestion d'organisation :



Nos fichiers aujourd'hui



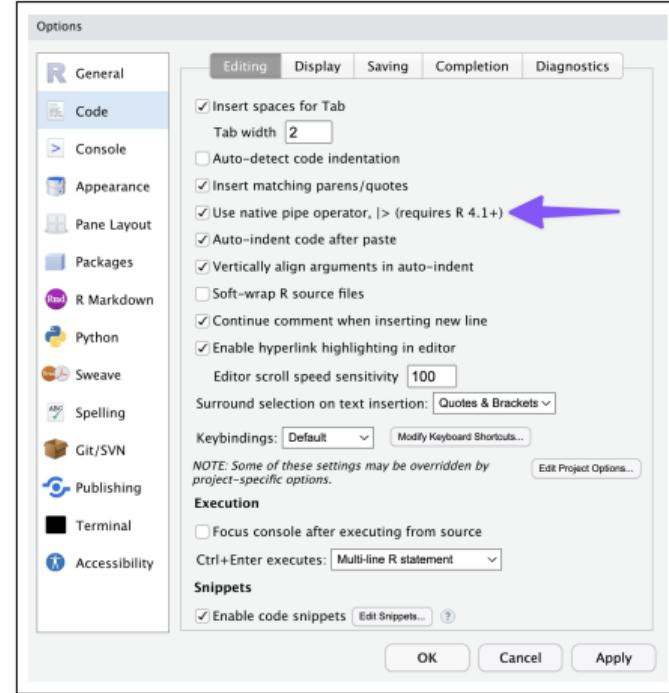
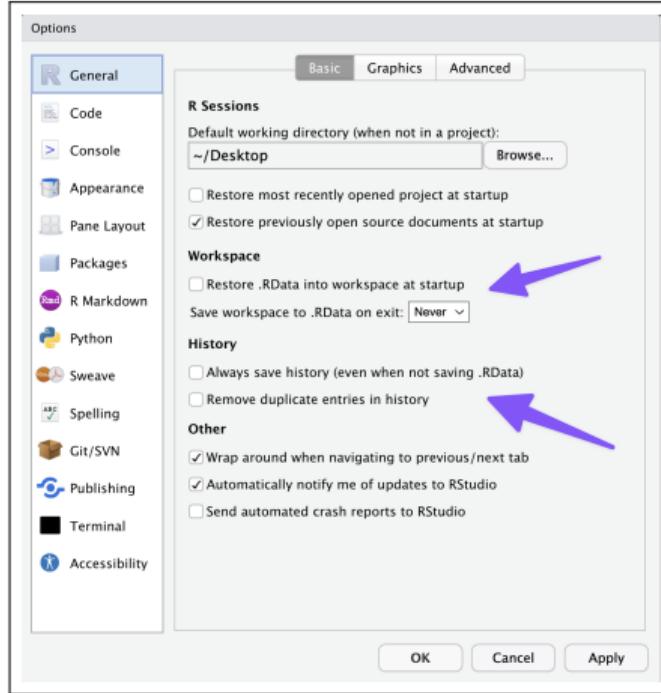
IMPORTANT

- Vous pouvez décider de ne pas suivre la suggestion d'organisation que je vous donne
You êtes libres de choisir votre propre organisation (ce n'est pas un cours d'informatique)
- **Dans ce cas-là :**
 - je comprends que **vous savez comment naviguer dans vos dossiers et vos fichiers**
 - ☞ je ne veux pas entendre la question « C'est où le fichier?! » pendant le semestre
- ☞ Si vous n'êtes pas capables de trouver vos fichiers, on sera bloqués



Quelques ajustements dans RStudio

Tools > Global Options...



On travaille dans RStudio maintenant



Synthèse : Les commandes et les fonctions d'aujourd'hui

```
1  read_csv(...)                      # importer un fichier csv
2  write_csv(...)                     # exporter un fichier csv
3  mean(...)                          # moyenne
4  sd(...)                            # écart-type
5  summarize(, .by = ...)           # créer un résumé avec n nouvelles colonnes
6  tribble(...)                      # créer un tableau (tibble) manuellement
7  glimpse(...)                      # visualiser les colonnes comme lignes
8                                # (idéal pour les tableaux avec plusieurs colonnes)
9  pivot_longer(names_to = "...",    # transformation « wide-to-long »
10            values_to = "...",
11            col = ....)
12 mutate(...)
13 filter(...)
14 select(...)

15
16 # Le « pipe » (|>) combine les fonctions d'une façon intuitive (raccourci Cmd/Ctrl + Shift + M) :
17 donnees |>
18   mutate(...) |>                  # créer des colonnes
19   filter(...) |>                # filtrer les données
20   select(...) |>              # sélectionner quelques colonnes
21   summarize(...)                 # créer un résumé
```

☞ Consulter les scripts et les solutions ici ↗



Références I

Barnier, J. (2023). *Introduction à R et au tidyverse*. Available at <https://juba.github.io/tidyverse/>.

