
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...?

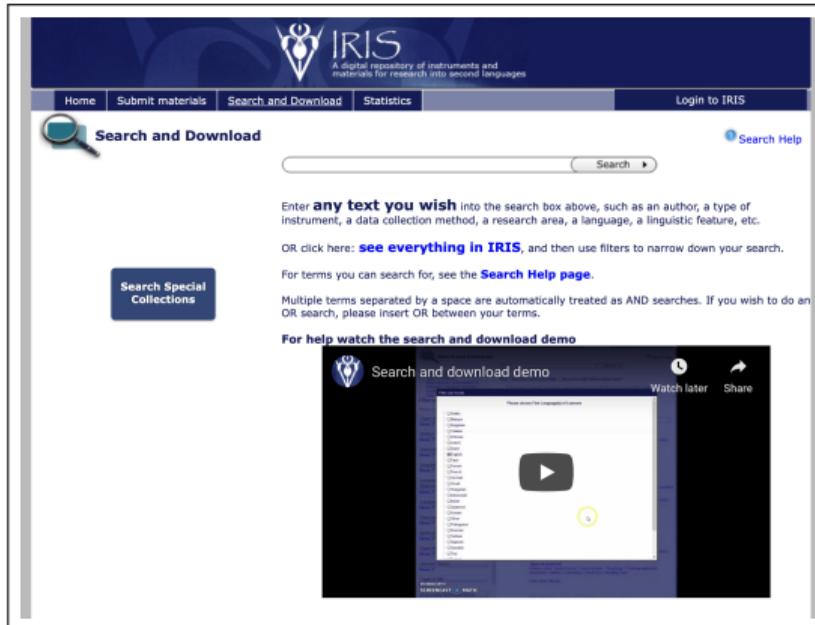
- Écrites?
- Orales?
- Jugement?
- etc.



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↗

The screenshot shows the CHILDES website homepage. At the top left is the CHILDES logo, which consists of two blue speech bubbles with the word "CHILDES" written vertically between them. To the right is the text "Child Language Data Exchange System". Below the header is a brief description: "CHILDES is the child language component of the [TalkBank](#) system. TalkBank is a system for sharing and studying conversational interactions." The main content is organized into several columns:

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

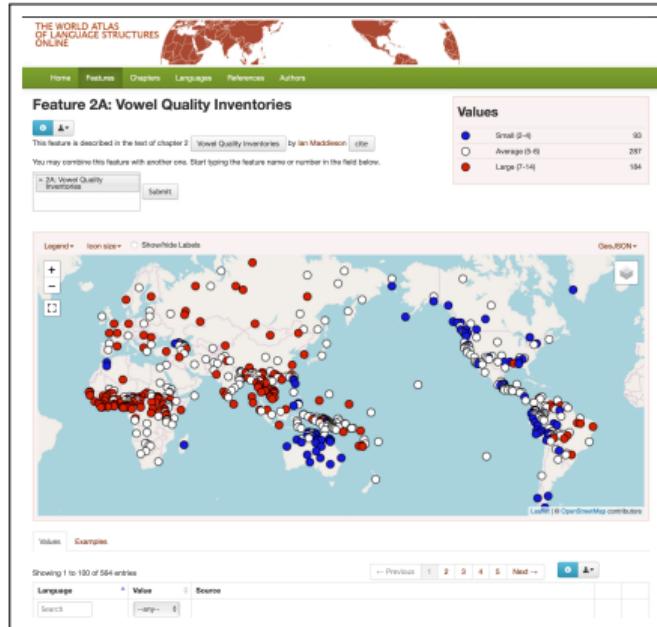
¹Une partie du projet [TalkBank](#)↗.



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 "earlobe". To its right is a stylized illustration of lips. To the right of these illustrations, the text "the speech *accent* archive" is written in a large, serif font. Below this, there is a vertical menu with the following options: "how to", "browse", "search", "resources", and "about". A small illustration of a brain is positioned between the ear and the lips. To the right of the menu, there 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 buttons for Facebook and Twitter are located 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

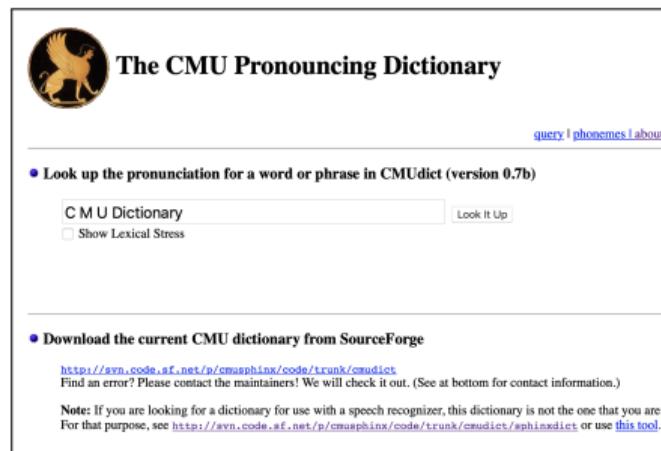
last updated: 15 january 2019 2780 samples



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:

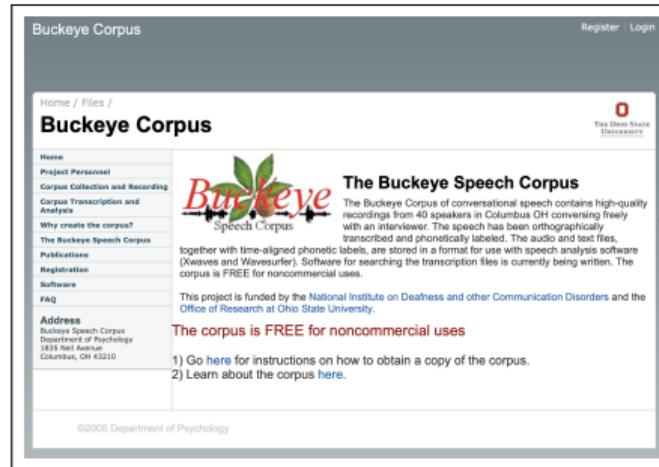
- Look up the pronunciation for a word or phrase in CMUDict (version 0.7b)**
A search input field containing "C M U Dictionary" with a "Look It Up" button next to it. There is also a checkbox labeled "Show Lexical Stress".
- Download the current CMU dictionary from SourceForge**
A link to the SourceForge 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", "Project Personnel", "Corpus Collection and Recording", "Corpus Transcription and Analysis", "Why create the corpus?", "The Buckeye Speech Corpus", "Publications", "Registration", "Software", and "FAQ". Below the navigation is a large logo for "Buckeye Speech Corpus" featuring a green leaf graphic and the word "Buckeye" in red. To the right of the logo, the text "The Buckeye Speech Corpus" is displayed. A detailed description follows, mentioning high-quality recordings from 40 speakers in Columbus OH, orthographic transcription, phonetic labeling, and audio/text files. It notes the corpus is FREE for noncommercial use and funded by the National Institute on Deafness and other Communication Disorders and the Office of Research at Ohio State University. At the bottom, there's a copyright notice for "©2005 Department of Psychology".

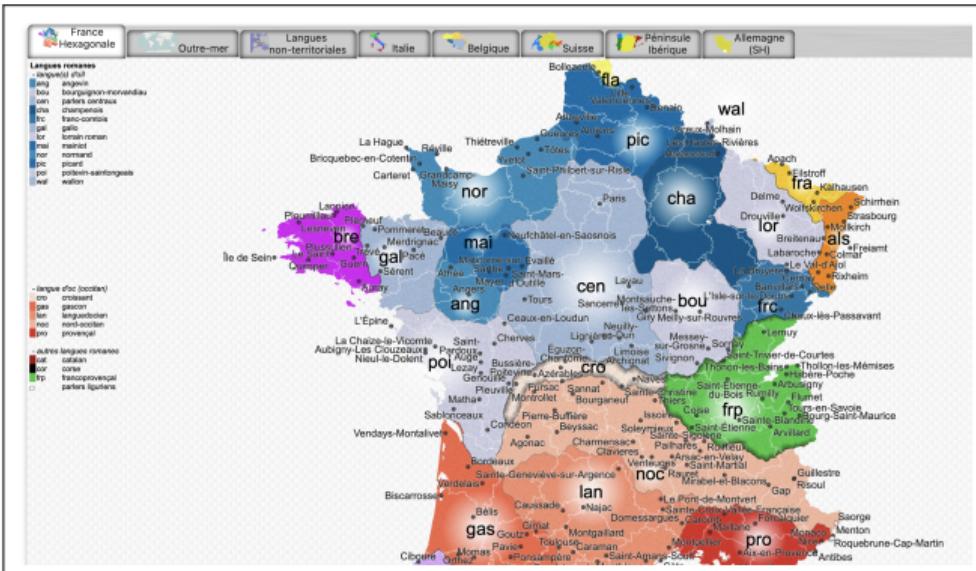
(Script pour échantillonner des mots à partir du corpus Buckeye)



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 cadrecl Yiil(version 1.1.22)



15 sur 21

Intro à R



Questionnaire

forms.office.com/r/XgXnS2Y8wD



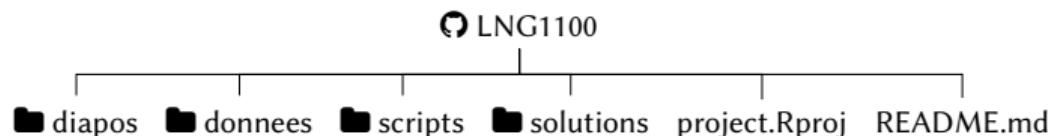
Projet R pour LNG-1100

Pourquoi un projet...?

- On concentre tous les fichiers du cours dans un seul dossier (dépôt Git)
- RStudio connaîtra déjà la localisation des fichiers à partir du fichier `project.Rproj`

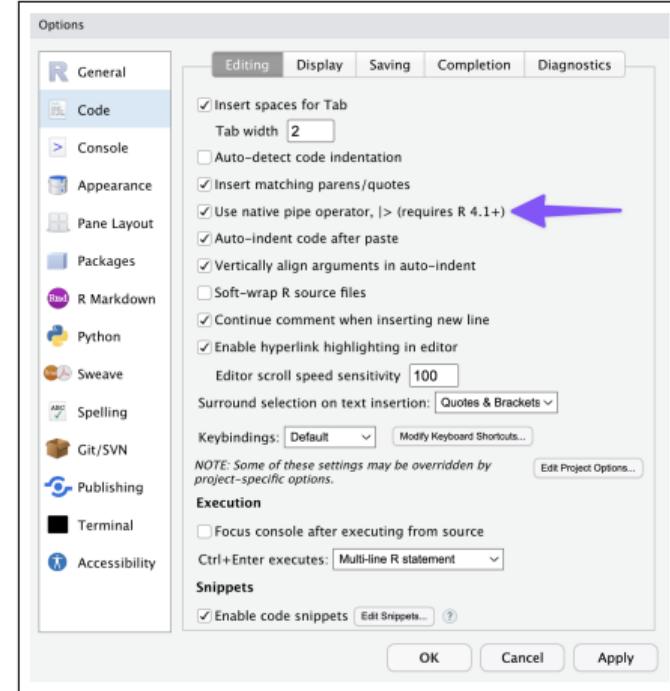
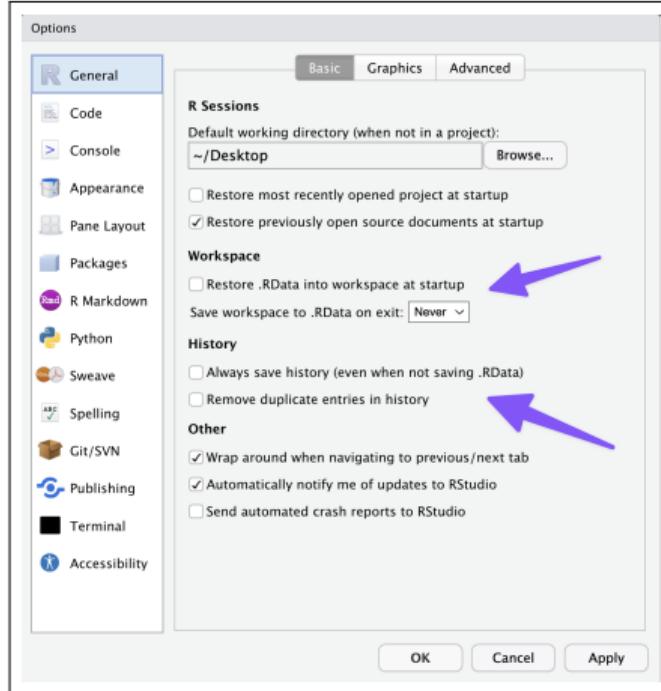
Maintenant, on continue sur RStudio (consultez les fichiers plus tard)

☞ Voici la structure de notre dépôt Git :



Quelques ajustements dans RStudio

Tools > Global Options...



RSTUDIO



Références I

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

