
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 (F0)^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?



Votre projet de recherche

- Un projet de recherche doit être **réaliste**, cela veut dire que :
 1. vous avez le **temps** pour suivre toutes les étapes nécessaires
 2. vous avez la **connaissance de base** exigée dans le projet
 3. vous connaissez les aspects **techniques** et **logistiques**¹ impliqués
- ☞ Un projet prend le temps, même si les points 1–3 sont très clairs
- **Deux problèmes typiques :**
 - on sous-estime le temps et la difficulté des étapes nécessaires
 - on est trop générale dans notre question ou dans notre sujet de recherche

¹Approbation éthique, collecte de données, paiement des participants, accès aux équipements, etc.



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...?

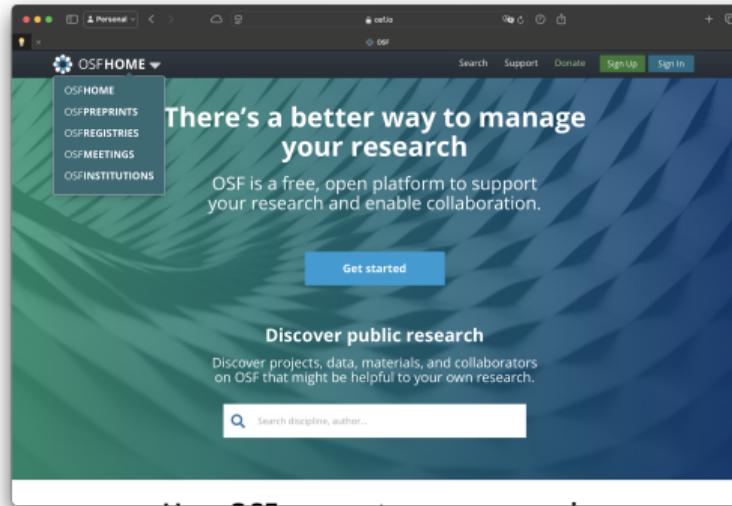
- Écrites? Facile à collecter, mais parfois difficile à analyser
- Orales? Comment coder les données?
- Jugement?
- etc.



Bases de données

OSF

- Base de **données** générale (plusieurs domaines/sujets) : [osf.io[↗]](https://osf.io)
- L'option la plus importante actuellement
- *preprints* + matériels supplémentaires (y compris des données, des *scripts*, etc.)



Bases de données

IRIS

- Base de données des études sur une langue seconde : iris-database.org ↗



Bases de données

CHILDES

Spécifique pour l'acquisition du langage² : childestalkbank.org↗

The screenshot shows the CHILDES website homepage. At the top left is the CHILDES logo, which consists of a stylized blue and white speech mark icon with the word "CHILDES" written vertically next to it. To the right is the "Child Language Data Exchange System" logo, featuring a small blue icon of a child's head with the word "child" written below it. The main content area has a light gray background with several sections:

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

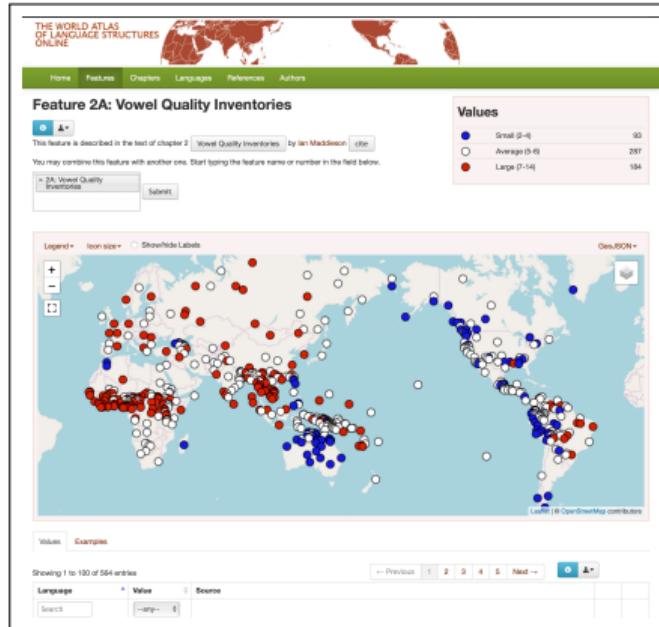
²Une partie du projet TalkBank↗.



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. A vertical menu bar on the left lists "how to", "browse", "search", "resources", and "about". The main title "the speech *accent* archive" is centered above a descriptive paragraph. Below the paragraph is a link to practice phonetic transcription. At the bottom are social media sharing buttons for Facebook and Twitter, and 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

Tweet

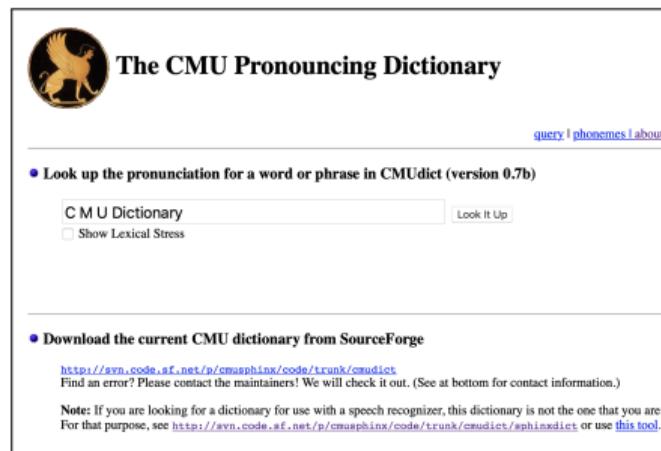
GEORGE MASON UNIVERSITY



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 contains two sections:

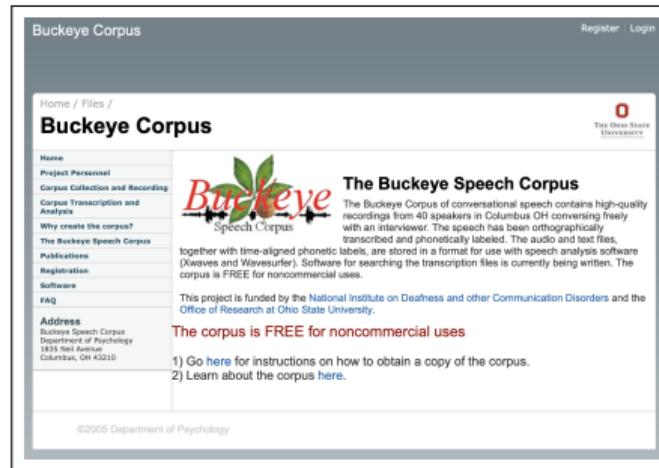
- Look up the pronunciation for a word or phrase in CMUDict (version 0.7b)**
A search input field containing "C M U Dictionary" with a "Look It Up" button next to it. There is also a checkbox labeled "Show Lexical Stress".
- Download the current CMU dictionary from SourceForge**
A link to the SourceForge 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.)"
Text 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 'Buckeye Corpus' on the left and 'Register | Login' on the right. Below the navigation is a header with the text 'The Buckeye Speech Corpus' and a logo featuring a green 'Buckeye' wordmark above the words 'Speech Corpus'. To the right of the logo is the Ohio State University seal. The main content area has a dark grey background. On the left, there's a sidebar with links like 'Home', 'Project Personnel', 'Corpus Collection and Recording', 'Corpus Transcription and Analysis', 'Why create the corpus?', 'The Buckeye Speech Corpus', 'Publications', 'Registration', 'Software', and 'FAQ'. In the center, there's a large text block about the corpus, mentioning it contains high-quality recordings from 40 speakers in Columbus OH conversing freely with an interviewer. It also notes that the speech has been orthographically transcribed and phonetically labeled, and that audio and text files together with time-aligned phonetic labels, are stored in a format for use with speech analysis software (Xwaves and Wavewerker). It states that the corpus is FREE for noncommercial uses. At the bottom of the page, there's a copyright notice: '©2005 Department of Psychology'.

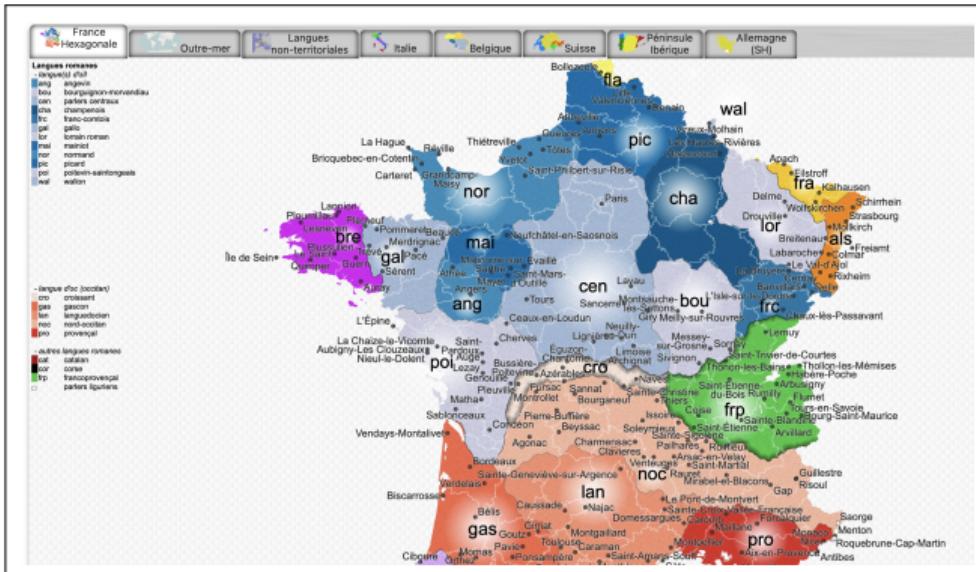
(Script pour échantillonner des mots à partir du corpus Buckeye)



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

Recherche sur la société et la culture Québec

Social Sciences and Humanities Research Council of Canada Conseil de recherches en sciences humaines du Canada 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 cadre YII (version 1.1.22)



Intro à R



Questionnaire

forms.office.com/r/XgXnS2Y8wD



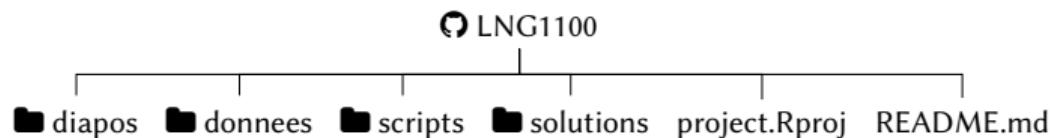
Projet R pour LNG-1100

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

☞ Voici la structure de notre dépôt Git :



RSTUDIO



Références I

Julien BARNIER : *Introduction à R et au tidyverse*. 2023. Available at <https://juba.github.io/tidyverse/>.

