Second Language Speech Data:

challenges in creating a corpus of English-L2 speech for open access

Ronald Lima Jr
ronaldojr@ufc.br
ronaldolimajr.github.io

Universidade Federal do Ceará

Importance of sharing data

Importance of sharing data

Replication crises

- reproducibility, replication, critique → development of the field
- *“with great results come great responsibility”*

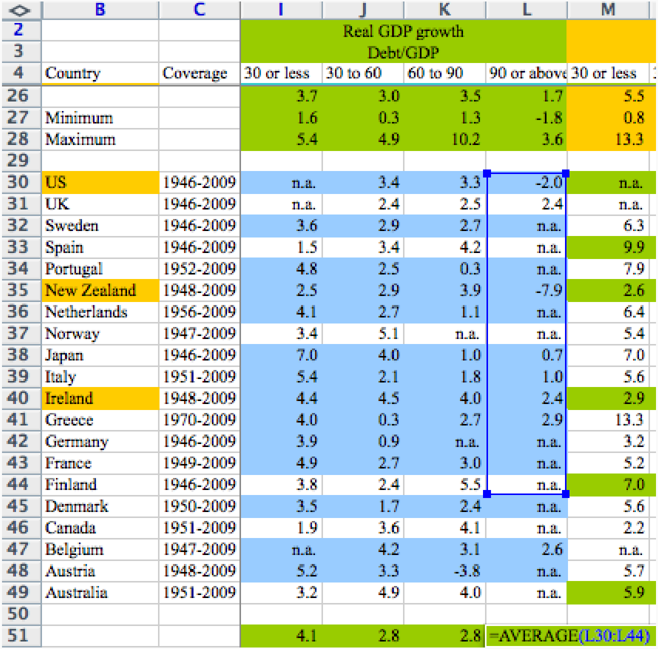
Importance of sharing data

Replication crises

- reproducibility, replication, critique → development of the field
- *"with great results come great responsibility"*
- Example 1 [McElreath 2020]: 2015, high-impact journal, 1170 children, negative association between religiosity and generosity
  - → countries (categorical) were entered as continuous
  - → Canada (country #2) was twice as much country as the US (country #1)
  - paper retracted, happy ending because data was shared

Importance of sharing data

- Example 2 [McElreath 2020]:
  - 2010: "Growth in a time of debt" by Reinhart & Rogoff
  - 2013: "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff" by Herndon, Ash & Pollin



◇	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

Importance of sharing data

Again, happy ending because data was shared (upon request), but...

what if data hadn't been shared?

what about those that are never shared?

Importance of sharing data

- When researchers are contacted to share data (data sets, scripts, procedures)
    - don’t reply
    - don’t have the data / all files / details anymore
    - files are disorganize; can’t remember which ones are the most recent, actually used in the paper

Importance of sharing data

- When researchers are contacted to share data (data sets, scripts, procedures)
  - don't reply
  - don't have the data / all files / details anymore
  - files are disorganize; can't remember which ones are the most recent, actually used in the paper
- Default should be:
  - keep files organized (names of files, folder system, version control)
  - keep them on-line, for back-up and to make them public

Importance of sharing data

- When researchers are contacted to share data (data sets, scripts, procedures)
  - don't reply
  - don't have the data / all files / details anymore
  - files are disorganize; can't remember which ones are the most recent, actually used in the paper
- Default should be:
  - keep files organized (names of files, folder system, version control)
  - keep them on-line, for back-up and to make them public

→ Include data sharing in your research project to be submitted to an ethics committee

Our project

Our project

Collect and publicly share longitudinal speech data of English-L2 by Brazilians

(Rosane Silveira, Ronaldo Lima Jr., Ubiratã Alves, Clerton Barboza)

- We're trying to do as we preach:
  - → Interab 12
  - → Silveira, R., R. M. Lima Jr., U. K. Alves & C. L. Barboza (2020) Construção de um banco de dados para pesquisas em fonética e fonologia de L2: um projeto interinstitucional, *Revista Colineares*, 7(1):81–113.
- Still needed: that researchers (users of our corpora) share their analysis files

Goal

- Longitudinal speech data of English-L2 by Brazilian learners

  → 4 different states (RS, SC, RN, CE)

  - undergrads of English Language Teaching +
    undergrads of Executive Secretariat +
    learners in a language course
  - Portuguese-L1 also recorded

Challenges & Possibilities

- Creating an instrument to collect as much as possible in a feasible manner

  → Proficiency in English

  image description + assessment of four judges

Challenges & Possibilities

- Creating an instrument to collect as much as possible in a feasible manner

  → Proficiency in English

  image description + assessment of four judges

  → 2 tests of production in Portuguese (vowels and consonants)

  "Em x *e* y *temos* z*"* & *"*x *e* y *têm uma/duas sílaba(s)"*

Challenges & Possibilities

- Creating an instrument to collect as much as possible in a feasible manner

  → Proficiency in English

  image description + assessment of four judges

  → 2 tests of production in Portuguese (vowels and consonants)

  "Em x e y temos z" & *"x e y têm uma/duas sílaba(s)"*

  → 3 tests of production in English (vowels, consonants, syllables)

  "I said x e y before I said z" & *"x and y don't sound like z"*

Challenges & Possibilities

Proficiency	**Image description** [30s, Likert, CEFR]
Production in Portuguese	**Test 1**: oral stressed vowels [70 words, 10 per vowel] **Test 2**: plosives and rhotics in onset; rhotics, lateral and nasals in coda; hetero-syllabic consonant clusters [68 words]
Production in English	**Test 1**: stressed vowels; plosives [38 words] **Test 2**: tautosyllabic consonant clusters; rhotics, laterals, nasals [39 words] **Test 3**: complex codas; dental fricatives; final -ed & -s [109 words]

Challenges & Possibilities

- Some data from SC and RN

  → CE and RS ready to start

Challenges & Possibilities

- Some data from SC and RN

  → CE and RS ready to start

- Keeping participants

  → Suggestions?

Challenges & Possibilities

- Some data from SC and RN

  → CE and RS ready to start

- Keeping participants

  → Suggestions?

- Infrastructure for high-quality recordings

  → soundproof booths in SC, RS and CE

  → high-quality recorders in silent rooms in RN

Challenges & Possibilities

- Pandemic

  → recordings continued in RN (on-line)

  → recordings paused in SC

  → CE and RS ready to start

Challenges & Possibilities

- Pandemic

  → recordings continued in RN (on-line)

  → recordings paused in SC

  → CE and RS ready to start

- Sharing the data (in a user-friendly interface)

  → grant (RNP, IBICT, CNPq) for data repositories

  (Juliana Soares Lima – Federal University of Ceará)

  → writing the Data Management Plan (DMP)

  → hopefully soon in dataverse

Changes/additions to be implemented

Data collection in CE and RS will still begin and could use changes/additions

- Collect data of students reading a text? Speaking freely?
- Collect production data of Portuguese-L1 every semester as well?
- Other?

Questions?

Suggestions?