A collaborative game-based approach to documenting linguistic variation in Brazil

Jenna Renjie Zhou

September 12, 2018

Abstract

We present a game with a purpose for the collection of speaker judgements relating to complementation alternations in Brazilian Portuguese. The game prompts users to rate sentence prompts in an environment themed around popular Brazilian funk musicians. The game may be played in a web browser, users can register for free or sign up with Facebook, and we have ethics approval for its deployment and for related data analysis. In this report we describe the decision-making and steps taken to develop the game, and discuss future work directions.

1 Introduction

Online data collection methods for linguistics are by now well established as a fast and cost-effective way to gather data from large numbers of speakers. One of the main benefits is the likelihood of collecting linguistic judgements from speakers of more varied demographics than traditional on-campus data collection from undergraduate students. This combination of broader demographics and larger participant numbers means that judgements hold more statistical power and are therefore more reliably representative of the truth. For instance, the 'Which English?' game[1] was used to collect grammaticality judgements from 669,498 online players, data which was subsequently used to analyse learning rates and the limits of ultimate attainment in second language learning (Hartshorne, Tenenbaum, & Pinker, 2018).

My aim was to design and develop a so-called Game With A Purpose (GWAP) to collect data on the use and acceptability of certain linguistic constructions in Brazilian Portuguese (BP), in a joint project with Dr Ioanna Sitaridou, Dr Andrew Caines (both of the University of Cambridge), and Professor Miriam Bouzouita of Ghent University. The sociolinguistic polarization of BP characterizes the language spoken in Brazil today, leading to a great deal of variation and macro-level sociolects (Lucchesi, 2009) – one linked to the formation of Cultured Portuguese (cultured norm) and another to Popular Portuguese (popular norm). These varieties reflect deep differences in the socio-economic environment in Brazil.

The popular norm includes different speech patterns such as Urban Portuguese, Rural Portuguese, and Afro-Brazilian Portuguese. Afro-Brazilian Portuguese is defined by a linguistic variety used in rural communities composed mostly of direct descendants of African slaves who settled in remote localities of the interior of Brazil. Some rural communities are remnants of *quilombos* ('hinterland settlements') and remain relatively isolated, thus making linguistic fieldwork more costly on all levels.

In contrast, an online game offers great potential for data collection and wide reach in modern society. Brazil, the world's fifth most populous country, is a Latin American champion of interactive entertainment. According to a survey conducted by SuperDataResearch, Brazil online games market is the fifth largest in the world and it accounts for 62% of the Latin American market. Counting almost 200 million people, Brazil's Internet penetration is growing at a rapid pace. Where the worldwide Internet penetration currently stands at 32%, Brazil's Internet penetration recently reached 50%, making the country a perfect market for digital games. According to Statista, Brazil's mobile user base is over 130m users and smartphone penetration is close to 50%.

[1] http://archive.gameswithwords.org/WhichEnglish

These statistics demonstrate how familiar Brazilians are with online games and that an online game may help us reach wide audiences for the purposes of studying sociolinguistic variation in Brazil.

Considerations in designing the GWAP include:

- The GWAP should be fun with a visually-appealing and intuitive user interface;
- The narrative of the GWAP is of utmost importance (Fort 2016), i.e. it needs an attractive theme (e.g. zombies in ZombiLingo, etc.) especially because we want to eliminate any implication that we want users to give 'correct' rather than natural answers (especially given the diglossic situation described above);
- It should incentivize participation and engagement with the task using gamification techniques (for example, showing leaderboards, achievements in a user profile, etc);
- User input on the examined linguistic phenomena is captured in a format appropriate for validation;
- User metadata (age, gender, geographic information, and so on) should also be captured in an ethically appropriate fashion (i.e with clearly informed consent).

The specific linguistic property we are interested in is BP complementation, since it is known to show a great deal of variation, with linguists often disagreeing over grammaticality judgments for phenomena such as the use of subjunctives and the distribution of inflected/personal infinitives (Sitaridou, 2014). For example –

> Então podem levarem um panito caseiro do forno, e comerem coentros. 'They could then take a home-made small bread from the oven and take corianders'
> Gostas?
> From one to three
> One: Gosto!
> Two: ok!
> Three: Não gusto não

For more example sentences, please refer to the Appendix.

The purpose of this UROP internship was to design and develop the GWAP according to the criteria listed above and for the purpose of collecting judgements on BP complementation. In future work we will add to the content of the GWAP, test and deploy the GWAP for data collection, and evaluate user responses. Further cycles of development and deployment may follow: the GWAP platform has been developed in a versatile way so that further mini-games may be incorporated in future to collect data on other linguistic phenomena. I now describe the progress made during the UROP internship, and the tasks which remain outstanding or of potential use before deployment of the GWAP. The app is available to use online at `http://jennazhou.pythonanywhere.com`

2 Web-app development

2.1 User Interface Design

Funk music is the dominant Brazilian pop music these days. Very prevalent and infectious, funk music can easily capture the public imagination. After discussion with Dr Ioanna Sitaridou and Dr Andrew Caines, I have decided that the questions for data collection will be presented in a format of interactive dialogue with the players, with a theme of funky music that is very popular among Brazilians.

The storyline of the game is that popular funk musicians invite the players to collaborate on writing lyrics for their new songs, with the sample sentences presented as lyrics in the speech bubble in the game to simulate a conversational dialogue with the players. The home page has a dark theme colour with funky musicians' figures as background to put the players into funky

music world. Every question page has a different theme colour and funky musician figure (which will be cartoon figure in the end due to the issue of personality rights), and a different format of presenting the question, to make the game more fun and appealing to the players. On the question page there is no button to return to the home page to discourage the players from quitting the game halfway, so that more complete datasets can be collected.

The player needs to register to play the game. After logging-in, a very simple public profile is displayed on the home page, showing the player's username, which is either the player's Facebook name or the partial email address that the player uses for registration. After starting the game, progress bar is shown on every question page to indicate the player's current progress. To create a sense of intensity and make the game more competitive and fun, a stopwatch is displayed and records the time already taken by the user so far in the game. Upon logging in, players can directly go from the home page to the leaderboard page to check on their ranking. Currently, the ranking system is based on the duration of completing all the questions by all players. The player can choose to delete account if the player no longer wants to keep any information.

After logging in to the account, the first-time players are required to fill in a data-collection form to indicate their age in years, their mother tongue language(s), country of birth and current country of residence. If the players are not playing the game for the first time, they will not be asked for the information any more and the game starts directly.

2.2 Back-end Structures

2.2.1 Server Set-Up

I chose to develop the app using the Flask framework, which is a micro web framework written in Python. Flask supports powerful extensions that can achieve what I need for the project. Additionally, Flask has well-supported free public host platforms which are safe in terms of data protection – one example is Python Anywhere, which I am using to host this website app for public access. As the data required to be stored is not complex nor in huge amounts so far, I am currently using an SQLite database engine to connect to database files from the Python program.

2.2.2 User Data Storage

As the main goal of the website app is to collect data from the players, the storage technique and data protection are essential. More details of data protection will be explained in the next section. This subsection explains the data storage technique used in this project.

A user database file is created for storing players' registration information, which is their email and password(hashed), and additional data that is considered as necessary data input for the research, such as the player's age, mother tongue language(s), country of birth and country of residence, as these are potential factors that may affect the player's grammatical judgements of the sentence structures. For those players who log in via Facebook, their email addresses and facebook names will be stored in the databases. The email address is used as the primary key to identify the player in the users' database table and store and extract the player's relevant information.

The player's inputs to all questions,together with the time duration the player takes to complete all questions, are only recorded after the player completes all questions at the end. The time durations of all players who have completed the whole game are then retrieved from the databases, ranked and displayed on the leaderboard.

When the player chooses to delete their account, all of the player's information stored in the database file is deleted, including the basic information and the input data provided by the player from playing the game.

2.2.3 Authentication System

There are two authentication systems presented to the players on the home page: registration and logging-in with email and password, and logging-in via their Facebook account which makes authentication easier and smoother for the players with Facebook accounts.

For the email registration system, only a valid email address and password are needed for registration. Hence, even without a Facebook account, players can also register for an account easily and conveniently. However, I still decided to integrate the Facebook login system into the game, because Facebook is very popular in Brazil and the number of Brazilian Facebook users is constantly increasing in the past few years.[2]. Having the Facebook login function smooths the login process for the players, reduces the friction in user experience, therefore potentially increases the number of players of the game.

It is not easy to integrate the Facebook login system with the email-registration system which was created first, as the overall login status is changed on the server side, while the Facebook login function makes changes on the client side. Therefore, I had to figure out a way to send a request to the server side after the response is returned to the client side from the Facebook API call when logging in via Facebook. I learned to do this by setting a flag for Facebook login status in local storage of the browser and sending a POST XmlHttpRequest when there is a change of the flag. Changes are then made on the server side when the request is received. A JSON file is used to send the Facebook login status data from the client side to the server side to notify changes in the connection status. Subsequently, a response containing current overall login status with login done in either way is received by the client side and changes are made accordingly in JavaScript.

2.2.4 Question Storage

Another separate database file is created to store the questions or stimulus and provided sample sentences I have received from Dr Ioanna Sitaridou. The primary key used in this database is the question number which is also the sequence of the questions showing up in the game. The questions or stimulus and the sample sentences will be retrieved from the database and presented on the question pages when the player starts playing the game.

2.3 Multi-player Support

As this is a website app game, there will be multiple players accessing the game simultaneously. Hence, a mechanism is needed to store the player's information and input data separately. Also, since a player's input data is only to be submitted for storage at the end of the game, a way of storing the player's input as the player continues to the next question page is needed to avoid losing data when making a new request.

Consequently, session and local storage are used in Python and HTML5 respectively to avoid one's input data overriding others' input data when multiple users are playing the game at the same time. When a player logs in to play the game, the player's email address which is the player's identification, the player's input data, and the player's final time taken to complete the game are all stored as session variables in a Python file and in local storage in HTML file. This not only ensures that all player's input data is remembered from request to request, but also avoids one player's data overriding another player's data when the two are playing the game simultaneously.

3 Ethics & data protection

We have ethics approval from the Department of Computer Science & Technology Ethics Committee (review #565). We described the project, listed the data we are collecting and made the following mitigating assurances relating to ethical data collection and data protection:

1. Explanation about the aims, personnel and use of data before informed consent for all participants is available at both home page and data collection form before starting the game properly. Participants have to acknowledge that they give their consent for the collection of the specific data required before playing the game for the first time;

[2] https://www.statista.com/statistics/244936/number-of-facebook-users-in-brazil/

2. Information for participants to contact project personnel is available at the home page. Participants can contact the project personnel via email shown under "Contact Us" , and functionality to self-manage data by deleting user account is also available under public profile displayed on the home page once participants sign in;

3. Secure data storage using Python Anywhere servers. Meanwhile, the password stored in the users' information database is encrypted using Passlib, which is a password hashing library in Python. When data downloaded for analysis by project staff, identifying information (email addresses / Facebook IDs) will not be included.

We have ethics approval for deployment of the app and collection of user data until 30 June 2019.

4 Future research & development

4.1 Graphics

Due to the above-mentioned personality rights problem, the background images of the Brazilian funk musicians will be changed from the photos of the real people to graphics of cartoon Brazilian funky figures. There may be a number of cartoon figures with different characteristics, doing different actions to add more variety to the game so that it is more fun. I will continue working on creating some sample graphic figures that fit the theme and style of this game.

4.2 Extent of Gamification

Currently the website app game is more like a questionnaire with graphic elements and animations to make the data collection process more fun and interactive. I suggest that mini games, such as identifying the singer of the funk music played in the game, can be added between question pages to increase the extent of gamification of the website app, so that the website app becomes more like a competitive online game with small sub-tasks for players to complete. Additionally, players can score points in the mini games, which can then be used to improve the ranking mechanism.

4.3 Scoring System

The ranking system is based on the time duration the player takes to complete all questions at the moment, which may not be the best approach for ranking the players. With the mini games added between question pages for players to score points, the total points scored and the time taken to complete the whole game by individual player can be input into a function to calculate the final overall score of the player, which can then be used for ranking. Consequently, the ranking system will be more applicable and effective in attracting more participants.

4.4 Language

Currently all text is in English. However, as the game is for Brazilians, the language used for the final version should be Brazilian Portuguese. Hence, translation of the text in the game will be needed.

4.5 Ethics and Data

Information about ethics and data regulations followed by the game is to be drafted and displayed when the players click on 'ethics and data' on the data collection form before starting the game properly to inform the players about details of the data collected and data protection regulations.

4.6 Contact Information

Contact information of relevant research personnel is to be displayed under 'Contact Us for More Information' after discussion of privacy issue, in case any professionals or players need to contact the project personnel for more information.

4.7 Introduction Video

An introduction video in Brazilian Portuguese is to be made and inserted under 'About the Game and Reward' to explain to the players how to play the game.

4.8 More Functionalities

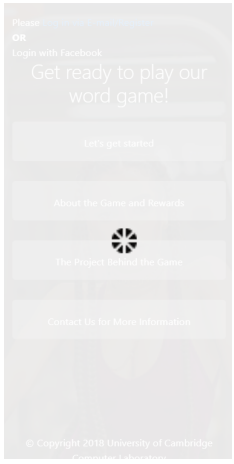
Suggested functionalities that can be added in the future:

- Although button to 'Delete Your Account' is added to HTML code, the functionality is not created yet. Code in python file is needed to make deleting account functional.
- If the scale of the game increases extensively and a variety of mini games is added to the app, a public profile may be created for individual player with the functionalities of changing username, changing password, setting public profile picture, checking highest scores obtained so far and receiving notifications of new update to the app.
- When there is a minimum amount of data available for analysis, machine learning techniques can be implemented to classify the future players according to the input supplied into different regions. If the classifier can be accurate enough, the main goal of the game can be modified to 'guess' which particular region of Brazil does the player come from.

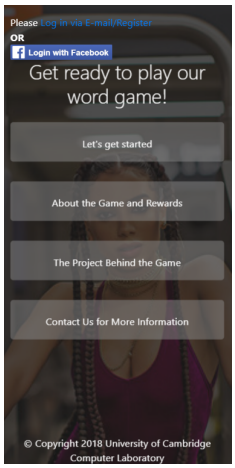
5 Screenshots

As it is presumably that most Brazilians have access to smart devices such as smart phones, and most people browse websites on mobile devices more frequently than desktops, I mainly designed the app using a viewport of mobile devices. Hence, I only included the screen shots of the app view in a mobile viewport.

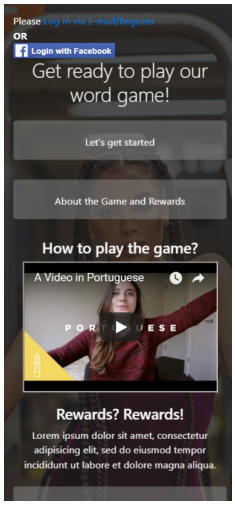
- A loading page is displayed when the browser is still in the process of connecting to the server and when HTML components have not been fully loaded yet to indicate the users that it is still loading.



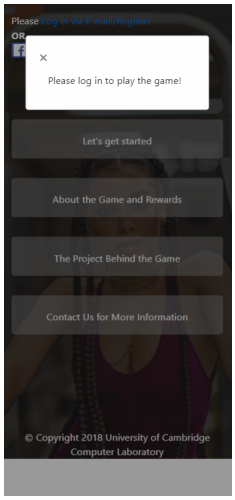
- This is the home page. Two authentication options are available when just opening the website.



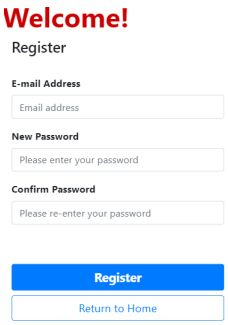
- An introduction video is available in Portuguese to explain how to play the game. The reward system is also explained in this section.



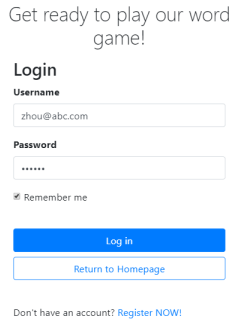
- When a player attempts to start the game without logging in, a notification pops up to remind the player to log in.



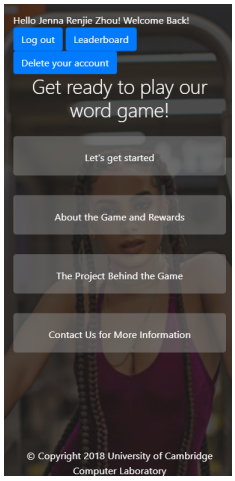
- For those players who do not have a Facebook account, they can register using an email address and password. Registration is fast and convenient.



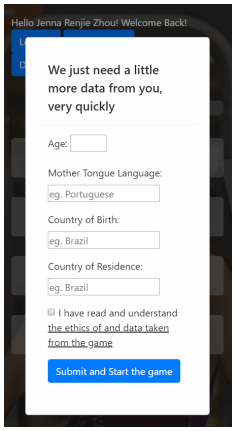
- To log in using email-password system. 'Remember me' may be ticked to remember the email address and the password to make it smoother and faster for players to login via this system.



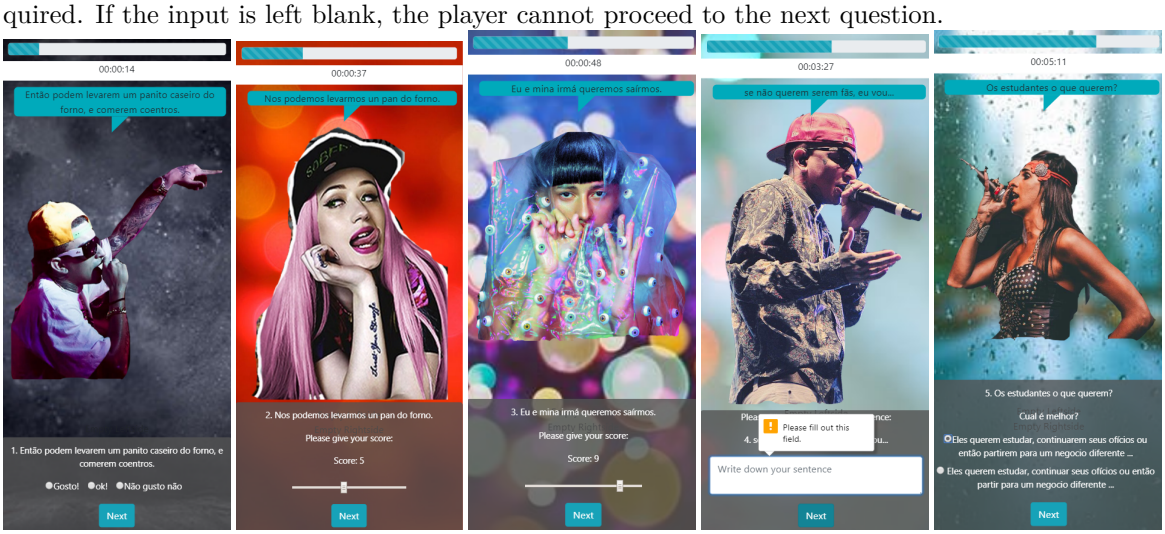
- The home page when the player successfully logs in to his or her account. In this case, the player logs in via the Facebook login system, therefore the player's Facebook name is displayed as the username on her public profile.

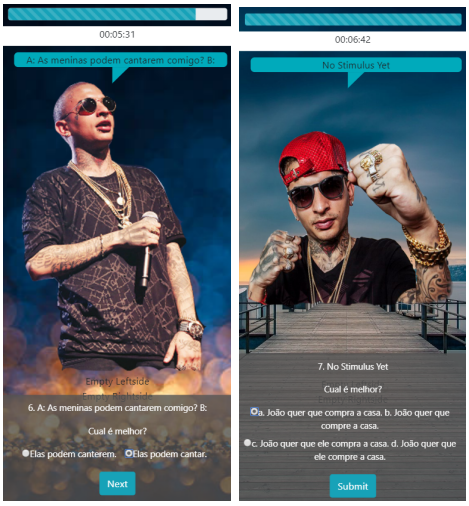


- Data collection form for collection of essential data from the first time players when they attempt to play the game. This form does not pop up for players who attempt to play the game from second time onwards.



- The following images are screenshots of questions pages. For every question, input is required. If the input is left blank, the player cannot proceed to the next question.





- Finally, after the player completes the game, the player can check his or her ranking on the leaderboard.

6 Acknowledgements

I would like to express my deepest appreciation to those who provided me significant help in my project. A special gratitude I give to Dr Paula Buttery, Dr Andrew Caines, Theodoros Michalareas, Felipe Schuery, and Ms Jane Walsh for their crucial guidance and assistance in the past 10 weeks.

References

Hartshorne, J., Tenenbaum, J., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263-277.

Lucchesi, D. (2009). História do contato entre línguas no brasil. In . I. R. D. Lucchesi A. Baxter (Ed.), *O português afro-brasileiro.* Salvador, Bahia, BR: Edufba.

Sitaridou, I. (2014). Modality, antiveridicality, and complementation: The Romeyka infinitive as a negative polarity item. *Lingua*, *148*, 118-146.

Appendix

Sample Questions

Dr Ioanna Sitaridou wants to test the following structures:

- 'can/want' + inflected infinitive (in coreference); prescriptive grammar says: plain infinitive only
- 'want' + indicative (in disjoint reference); prescriptive grammar says: subjunctive only

Data for Brazilian Portuguese GWAP

- Então podem levarem um panito caseiro do forno, e comerem coentros.
  'They could then take a home-made small bread from the oven and take corianders'
  Gostas?
  From one to three
  One: Gosto!
  Two: ok!
  Three: Não gusto não
- se não querem serem fãs, eu vou FILL IN (perhaps it can bleep or sth invitin them to fill it in) 'if they don't want to be fans, I will . . . '

  Gostas a frase?
  One: Gosto!
  Two: ok!
  Three: Não gusto não
- Stimulus: in a bubble: Os estudantes o que querem?
  Eles querem estudar, continuarem seus ofícios ou então partirem para um negocio diferente ...
  'They wanted to study ... or then leave to a different business ...

  Eles querem estudar, continuar seus ofícios ou então partir para um negocio diferente ...
  'They wanted to study ... or then leave to a different business ...

  Cual é melhor?
- Stimulus:
  One character says: As meninas podem cantarem comigo?
  The other character says:
  - Elas podem canterem.
  - Elas podem cantar.

  Tick your favourite answer.
- No stimulus yet. This is a scoring question:
  - Nos podemos levarmos un pan do forno.('We can take a bread from the bakery.')

  Please slide to score the sentence on a scale of 10.
- No stimulus yet. This is a scoring question:
  - Eu e mina irmá queremos saírmos.('My sister and I want to leave.')

  Please slide to score the sentence on a scale of 10.
- No stimulus yet. This is a binary choice question:
  - a. João quer que compra a casa.('João wants that pro buy.IND.3SG the house')
    b. João quer que compre a casa.('João wants that pro buy.SUBJ.3SG the house')

– c. João quer que ele compra a casa.(‘João wants that hej buy.IND.3SG the house’)
  d. João quer que ele compre a casa.(‘João wants that hej buy.SUBJ.3SG the house’)

Tick your favourite answer.