Predicting Errors with Second Language Acquisition Modeling

Denis Kapelyushnik*

Abstract

In 2018, a challenge on Second Language Acquisition Modeling was organised by Duolingo AI in conjunction with the 13th BEA Workshop and NAACL-HLT 2018 conference. One of the key findings of the challenge was the fact that a choice of a learning algorithm (for the task) appears to be more important than clever feature engineering. This research paper for the Linguistic Data: Quantitative Analysis and Visualisation course is aimed to explore if any available or synthesised feature can be used to predict potential errors. The Null Hypothesis Significance Testing framework will be used for analysis.

1. Metadata

The dataset used for this paper comes from B. Settles et al. (2018). To 7M words produced by more than 6k learners of English, Spanish, and French using Duolingo, an online language-learning app, were collected for the Second Language Acquisition Modeling (SLAM) task. The more detailed task description and results achieved by contestants are available on the official task page[1].

Only `train` splits prepared by Burr Settles (2018) were used in this project. A dataset per language pair was split into two files[2], e.g. `fr_en_metadata.csv` and `fr_en_sessions.csv`.

The data for this task are organized into language pairs: `es_en` — Spanish learners (who already speak English), and `fr_en` — French learners (who already speak English). The `en_es` part — English learners (who already speak Spanish) - is not included into this project,

Both `*_metadata.csv` and `*_sessions.csv` contain data separated by tabs (no headers):

Table 1: Content of the *_metada.csv files

Column name	Description
user_id	generated during data anonimisation
country	a 2-character country code
days	day of usage (a double)
client	android, ios, or web
session	lesson, practice, or test
format	reverse_translate, reverse_tap, or listen
time	duration of the answer in seconds
session_id	use it to join metadata and sessions
n_tokens	a number of tokens used in the task
n_errors	a number of errors a user made
prompt	prompt (no prompt in listening)

*HSE University, dmkapelyushnik@edu.hse.ru

[1]http://sharedtask.duolingo.com/2018

[2]To reproduce this paper, follow the instructions specified in the data folder of the project github repository.

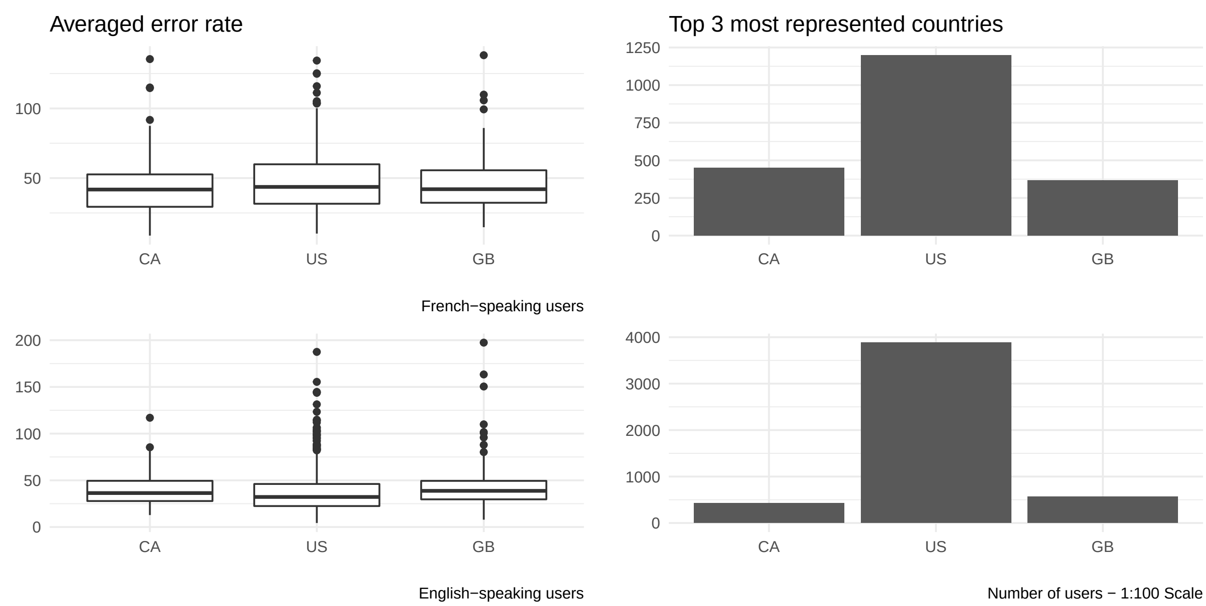
Table 2: Content of the *_sessions.csv files

Column name	Description
session_id	unique ID for a sesssion
task_token_id	location of a token in a task
token	word
POS	part of speech in UD format
morph	morphological features in UD format
ud_edge_label	dependency edge label in UD format
ud_edge_head	dependency edge head in UD format
label	to be predicted (0 or 1)

2. Descriptive statistics

Countries

Overall, there more than 100 locations where people use the app. As the number of users can differ significantly, it was decided to limit the number of countries - only USA, Canada and Great Britain are used for this project.



Tasks

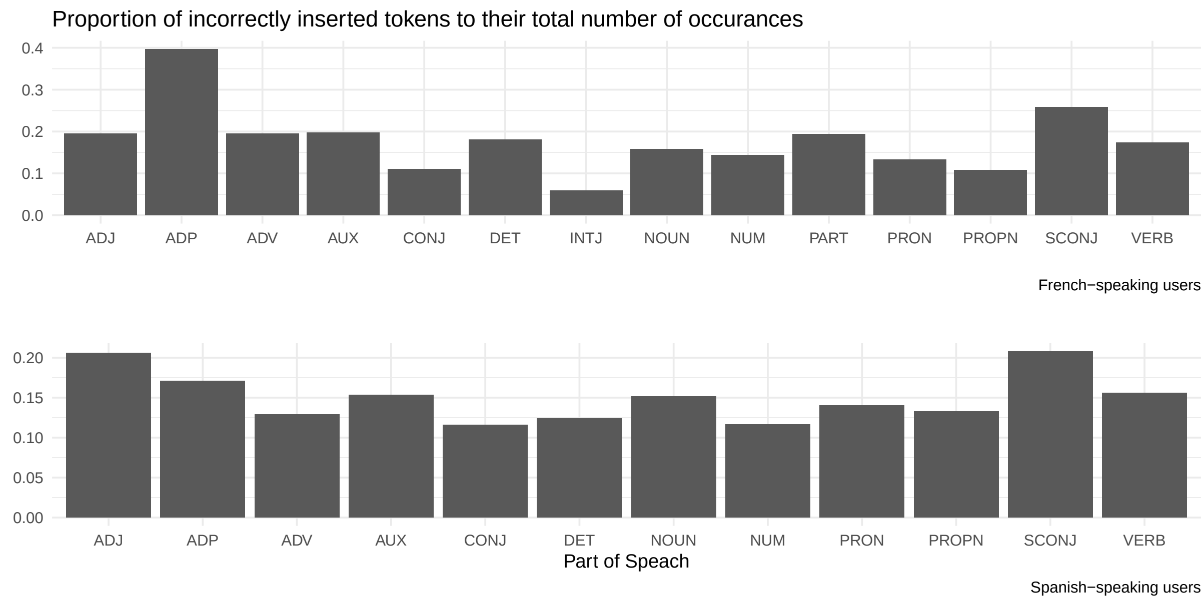
The data is collected for a 30-day period, during which users engaged in different formats of the tasks, namely `listen` (listen and translate into the source language), `reverse_tap` (drag tokens in a correct order) and `reverse_translate` (read and translate into the source language). Only `listen` and `reverse_translate` tasks require typing thus they are more prone to errors[3].

[3]Indeed, minimal edit distance is used to handle mistyping but it depends on a token, e.g. *you* will not be accepted for *your* even if edit distance is only 1

Table 3: Average Error Rate

Task Type	Spanish	French
listen	0.3942646	0.6736502
reverse_translate	0.6871858	0.6922757
reverse_tap	0.1383058	0.2081322

A task can contain from 1 to 14 tokens (depending on the language). Each token has a set of features assigned to it by B. Settles et al. (2018) using the Google SyntaxNet dependency parser and the language-agnostic Universal Dependencies tagset[4].



In the `fr_en` dataset, `PUNCT` and `X` tokens have the largest error_rate. It was excluded from the graph as the former refers to `-` in constructions like `Qui sont-ils?` and the latter mostly to `t` in consruction like `Qu'a-t-il?`. In both cases, error refers to the `-` character in a question. In general, the app does not penalise users for absence of punctuation marks, so they may just skip it thus "making a mistake".

Table 4: Distribution of Errors Tagged as PUNCT by Task Format

Format	Number of Errors
reverse_translate	992
listen	375
reverse_tap	1158

Another option is that this mistakes refers to word order issues. Some evidence for it comes from the fact these mistakes are mostly happen in `reverse_tap` tasks, which does not assume any typing at all. It is difficult to decide if it is a user-interface issue or a `real` mistake without seeing the actual user input thus no further exploration is possible.

The Spanish part of the graph does no include `SYM` and `X` tokens as the former occurs only once and the latter - a `Sí` token which is rarely put incorrectly.

[4]Parse errors may occur.

Both datasets have two distinct groups of tokens with a largest number of errors. For the French-English language pair, they are `ADP` and `SCONJ`. For the Spanish-English language pair, they are `SCONJ` and `ADJ`. In the next part of the project, I will explore if any particular feature may help to predict an error during SLA.

3. Exploring most common mistakes

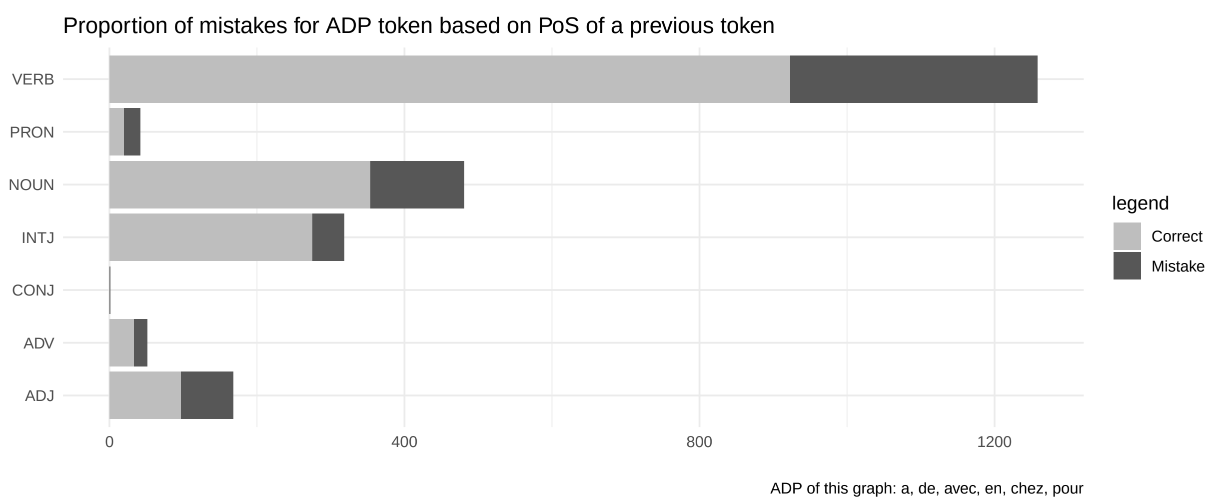
French-speaking learners

The most “erroneous” token with the `ADP` tag is a preposition `de`. There are three variants of its spelling in Top 5 most common mistakes.

Table 5: Most common mistakes for the ADP tag

Token	Quantity
D’	2249
de	379
a	307
De	150
avec	114
en	104
d’	99
Tu	81
Dans	80
chez	58
pour	58

It is quite unexpected, the two out of three variants refer to the the beginning of the sentence: `De` and `D'`. Apparently, it refers more to some labeling artifacts or a simple carelessness and does not deserve much attention. In these case, only `de`, `a`, `avec`, `en`, `chez`, `pour` will be explored more in-depth.



There is a significant share of errors if a preposition goes after `ADJ`, `VERB` and `NOUN`. It might be useful to use features of these tokens to predict with logistic regression. All of them share the `Gender` features, so let’s test if these feature correlates with errors.

1. Chi^2 test for `Gender` feature in `ADJ` indicates that that these two variables are independent as the p-value is much higher than 0.05. We can safely accept the null hypothesis.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 0.2814, df = 1, p-value = 0.5958

## Bayes factor analysis
## --------------
## [1] Non-indep. (a=1) : 0.5970475 ±0%
##
## Against denominator:
##   Null, independence, a = 1
## ---
## Bayes factor type: BFcontingencyTable, poisson
```

To be on the safe side, I used the Bayesian framework for confirmation. The result is the same - the odds for the alternative hypothesis against the null are about 0.59:1.


```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 0.2814, df = 1, p-value = 0.5958

## Bayes factor analysis
## --------------
## [1] Non-indep. (a=1) : 0.5970475 ±0%
##
## Against denominator:
##   Null, independence, a = 1
## ---
## Bayes factor type: BFcontingencyTable, poisson
```


2. The following are the results of the Chi^2 test for `Gender` feature in `VERB`

indicates that that these two variables are independent as the p-value is much higher than 0.05. We can safely accept the null hypothesis.

Let's analyse if any particular feature The following script is used to separate morphological features in the `morph` column:

Errors

There are much more samples with 0 errors than with any number of mistakes altogether. Below its visualisation, regular and scaled using log10.

The number of errors is connected with the length of a task - there are even samples where all tokens were inserted incorrectly. The more interesting to know if any of the task formats is more difficult than the other.

```
# es_en_md %>%
#   select(format, n_errors) %>%
#   group_by(format) %>%
#   summary()
```

Very moderate positive correlation coefficients and a very small p-value are observed, so we can reject a null hypothesis and safely assume that there is a higher chance to make a mistakes in longer sentences. The plot is two visualize it.

Rejection

References

Settles, B., C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani. 2018. "Second Language Acquisition Modeling." In *Proceedings of the Naacl-Hlt Workshop on Innovative Use of Nlp for Building Educational Applications (Bea)*. ACL. https://doi.org/10.7910/DVN/8SWHNO.

Settles, Burr. 2018. "Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)." Harvard Dataverse. https://doi.org/10.7910/DVN/8SWHNO.