Data Analysis 2 Term Project

Gyongyver Kamenar (2103380)

Introduction

This paper analyze the differences of English correctness between people with several different native languages. Nations have different linguistic characteristics, cultural and social norms, and education systems as well. Therefore, people with different nationalities or native languages have differences in English learning. That's why, the motivation of this study is to see, that whether there are differences in English correctness between English learners of different native languages controlled for other relevant factors. It is a relevant question for everyone, because if there are significant differences, scientist can further investigate the main reasons behind it. It might have policy implications on education systems or recommendation of learning practices, methods and habits for English learners.

Data

Origin of the data

The used dataset was originally collected by Hartshorne et. al. (2018) who studied the critical period for second language acquisition. He collected data from 669,498 respondents through online English quizes. The ca. 10 minute-long English grammar quiz was deliberately developed to enabled the researcher to measure the respondent's syntactic English knowledge and also provided several demographic variables. The critical items of the quiz are the diagnostics of proficiency and there were additional items to distinguish between dialects. In my analysis I used the percentage of correct critical items as a proxy of English correctness.

- **English correctness** : percentage of critical quiz items correct

Further information about the data is available (here)[https://osf.io/pyb8s/] and in Hartshorne et. al. (2018).

Explanatory variables

In this subsection I list the available variables in the dataset that I found relevant explanatory variable. In some cases I did not use a given variable in my analysis and I explain the reason.

- **Age** : age of subject (numeric)
- **Gender** : gender of the subject, either male, female or other (categorical)
- **Native language** : subject's native language(s) (categorical)
- **Native English** : English is the subjects native language (binary)
- **Primary language** : subject's primary language(s) now (categorical)
- **Primary English** : subject's primary language is English (binary)
- **Psychiatric** : subject reported any psychiatric disorders (binary)
- **Starting age of English learning** : age at start of English learning (numerical)
- **Language status** : either monoeng (native speaker of English only), bileng (native speaker of English + at least one other lang), immersion learner(spending at least 90% of their life since age of first exposure in an English-speaking country), or non-immersion learner(spent at most 10% of post-exposure life in an English-speaking country and no more than 1 year in total) (categorical)
- **Years of English "learning"** : age - starting age of English learning (numerical)
- **Education** : Highest level of education (categorical)

Variables I found relevant but did not use:

- **Dyslexia** : subject reported difficulty with reading (binary)

Reason: there was no variation in this variable (no one with dyslexia)

- **Live with English** : subject lives with any native speakers (binary)

Reason: majority of the data were NULL values indicating missing observation and I cannot be sure what it means

- **Countries** : countries subject lived in (categorical)

Reason: too much distinct categories (thousands) with just few subjects, moreover most subject lived in 2 or more countries

- **Current country**: country currently lived in (categorical)

Reason: too much distinct values (hundreds) and NAs

- **English country years**: number of years living in English speaking countries

Reason: most observations are missing and a similar feature is captured by Language status variable

Scope

The data includes bilinguals and immersion learners as well, which is beyond the scope of my research question. The study's interests are subject with 1 native language (not bilingual) and in case it is not English, we want to investigate subject who are non-immersion learners (spent at most 10% of post-exposure life in an English-speaking country and no more than 1 year in total). Thus the language status is filtered for monoenglish or non-immersion.

Furthermore, I analyze subject with 1 native language and there are several different categories in the dataset. I decided to analyze just the top 9 native languages beside English, based on the number of subject in the dataset (top 1 is English), namely; Finnish, Turkish, German, Hungarian,Russian, Dutch, Polish, Swedish and Spanish in the respective order. Even from the least frequent native language Spanish there were more then 9200 subjects in the filtered dataset.

Age is also a relevant factor in English correctness, so subjects at least 14 years old to make they are able to learn a language.

Descriptive statistics

Please see below the descriptive statistics of the numerical variables.

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P25	P75	P95
Age	28.91	27.00	9.20	14.00	89.00	18.00	23.00	33.00	48.00
Starting age of English learning	9.82	10.00	3.85	1.00	69.00	5.00	8.00	11.00	16.00
Psychiatric disorder	0.03	0.00	0.18	0	1	0.00	0.00	0.00	0.00
English correctness	0.91	0.93	0.07	0.31	1.00	0.77	0.87	0.96	0.99
Years of English learning	19.09	17.00	9.06	1.00	80.00	8.00	13.00	23.00	37.00
Log(English correctness)	2.53	2.47	0.90	−0.81	5.25	1.18	1.90	3.01	4.14

The number of observations is 168561 for all of our key variables. The age of subjects is filtered to be at least 14 years, because below that they might not able to learn a language properly.

We can see, that the SD of English correctness variable is really small, the mean is 0.91 and the median is 0.93 so the distribution is left skewed. See the distributions Figure 2 . It might worth to use the log of English correctness, however it complicates the interpretation of the coefficients because that the correctness is already measured in percentage.

Exploratory data analysis

FIrstly, I checked the correlation between the non-categorical variables, see the correlation matrix in Figure 1. The English starting age has negative correlation with English correctness -0.21 and years of English learning -0.18 (because the latter is age - English starting age). Age and years of English learning have positive correlation with English correctness. Psychiatric disorder is not really correlated with anything. In the modelling phase, I will check this matrix to decide which variable to add or not to add.

As the focus is English correctness for different native languages, the Figure 4 shows the boxplot for each native language. The differences in English correctness are visible on the boxplot.

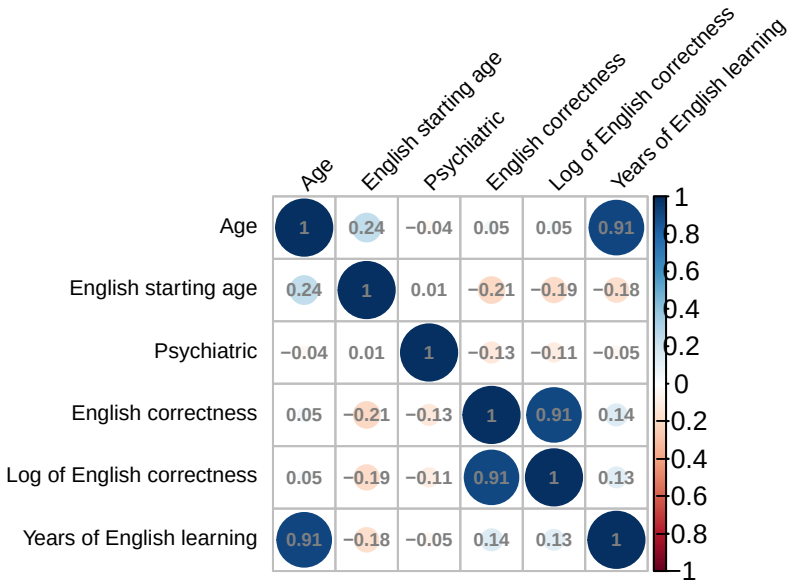

Figure 1: Correlation matrix

I also checked, whether there is a pattern between starting age of English learning and English correctness, as well as between age and English correctness. I plotted a non-parametric regregression (smoothed method) for all native languages as you can see on Figure 6 and Figure 7 . These can help me decide, whether to add any splines of interaction terms in the regressions. The pattern og age in English correctness seems pretty constant for all native languages, it some cases it has a slight downward slope but no knit points. However, the pattern of strating age of English learning shows steeper downward slope in the non-parametric regressions and in some cases the pattern change at around year 10. I also created a binned scatterplot for starting age of English learning on Figure 5 and the points change pattern exactly at 10 years. Probably adding a spline there can improve the model.

Models

Firstly, I estimeted a regression just with the native language categories and it already explained more than nearly 20% of the variance of log(English correctness). Then in the second model I added the relevant explanatory variables detailed before, so I added age, education, gender, psychiatric disorder and years of English learning to control for these variables. Because years of English learning highly correlates with age (it is calculated from that) in the third model is used the starting age of English learning to reduce multicollinearity issue.

In the fourth regression I added splines with a knot point at 10 years as suggested by the binned scatterplot in Figure 5 . Finally, in the fifth model I also added a knot point at 20 years also based on Figure 5 and Figure 6 . I was thinking about possible interaction of native language with a numerical variable based on the smoothed regressions, however, it would probably result overfitting. For example the pattern of Finnish Polish and Dutch cases on Figure 6 . See my estimation results in Table 2 .

Results

As you can see, all of the variables are significant in all models, except the Other gender probably due to it's low number of observations.
Based on the coefficienty, reasonability, Bayesian Information Criteria and Log Likelihood, my preferred model is the fifth :

$$\text{English correctness} = 0.86 + 0.083\ NativeDutch + 0.065\ NativeFinnish + 0.067\ NativeGerman + ... + \delta Z$$

where Z are standing for the controls, which includes controlling for age, gender, education and starting age of English learning. The base category of the native language is Turkish. From this model we can infer:

- when every covariates are zero, people expected to have English correctness of 0.86 (so it's the base category, Turkish people)
- when a subject's native language is Dutch, she/he has on average 0.083 higher English correctness (meaning 8.3 %) compared to the base Turkish, controlling for the other variables.
- so on for the other native languages ...

The highest language coefficient is of Native Dutch, followed by Native Swedish, German an Finnish. I have the t-statistic as 150.55 and the p-value as 0 in case of Native Dutch, which confirms my conclusion.

We know that the coefficients are statistically different from zero, so there is statistical difference compared to the base category (Turkish). However, I have to test whether the coefficients of the other native languages are different from each others. To show that, I have run a two-sided hypothesis test between all native language pairs x and y:

$$H_0 := \beta_x - \beta_y = 0$$
$$H_A := \beta_x - \beta_y \neq 0$$

There we 2 cases when I could not reject the null hypothesis at 1% significance level. The difference between the coefficients of native Hungarian and native Spanish anf between native Polish and native Spanish are not significantly different from zero.

Robustness check / 'Heterogeneity analysis'

The t-test statistics and the p-values in the estimation table confirms that the coefficients are different from zero. Partly, this is due to the low standard errors which are due to the high number of observations and the variance in the explanatory variables.

Throughout the models, the coefficient estimates are pretty stable, in some cases the coefficient in model 1 and model 5 is the same eg. Native Hungarian is 0.03 and 0.031 in simple model.

The quize about English correctness was carried out in 2014, and the world, the globalization and the popularity of English language changed since then. This suggest that the external validity is lower, then a test carried out recently. The countries and nationalities still have the most features than before, so those norms are mostly similar still now. The high number of observation, and the clode to random selection make us infer that the data and analysis has high external validity.

The biggest drawbacks of the research is the measurement of English correctness, and a proper measure or proxy for English learning time and intensity.

Conclusion

In this study, I was interested in whether there are differences in English correctness between people with different native languages. I analyzed subject with 9 different (and not English) native languages and controlled for other demographic variables like age, gender, education, starting age of English learning and psychological disorder. I used dummy variables to test the differences. The based category was Turkish, such people has 0.86 English correctness if every other variable is zero. I found, that there are indeed significant differences in English correctness between native languages controll for the demographic variables. In the cases, namely Hungarian - Spanish and Polish-Spanish the difference is not different from zero, however in every other case it is. The native language with the highest English correctness on average, controlled for the demographic variables, are Dutch, Swedish and German. This finding can have implications on education systems or language learning practices as well.

Appendix

Table 2: Estimated models

	Simple	Multiple	Multiple with starting age	Splines	Two knot splines
Intercept	0.862***	0.874***	0.874***	0.863***	0.862***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Native Dutch	0.087***	0.083***	0.083***	0.083***	0.083***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Native Finnish	0.072***	0.066***	0.066***	0.066***	0.065***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Native German	0.073***	0.069***	0.069***	0.068***	0.067***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Native Hungarian	0.031***	0.029***	0.029***	0.030***	0.030***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Native Polish	0.028***	0.026***	0.026***	0.027***	0.027***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Native Russian	0.019***	0.014***	0.014***	0.015***	0.015***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Native Spanish	0.036***	0.026***	0.026***	0.029***	0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Native Swedish	0.081***	0.075***	0.075***	0.075***	0.075***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age		−0.003***	0.000***	0.000***	0.000***
		(0.000)	(0.000)	(0.000)	(0.000)
Male		−0.010***	−0.010***	−0.010***	−0.010***
		(0.000)	(0.000)	(0.000)	(0.000)
Other gender		0.002	0.002	0.003	0.003
		(0.002)	(0.002)	(0.002)	(0.002)
Graduate Degree		0.032***	0.032***	0.032***	0.032***
		(0.001)	(0.001)	(0.001)	(0.001)
HighSchool Degree (12-13years ed)		0.017***	0.017***	0.017***	0.017***
		(0.001)	(0.001)	(0.001)	(0.001)
Some Graduate School		0.016***	0.016***	0.015***	0.015***
		(0.001)	(0.001)	(0.001)	(0.001)
Some Undergrad (highered)		0.025***	0.025***	0.024***	0.024***
		(0.001)	(0.001)	(0.001)	(0.001)
Undergraduate Degree (3-5years highered)		0.031***	0.031***	0.030***	0.031***
		(0.001)	(0.001)	(0.001)	(0.001)
Psychiatric		−0.045***	−0.045***	−0.045***	−0.045***
		(0.001)	(0.001)	(0.001)	(0.001)
Years of English learning		0.003***			
		(0.000)			
Starting age of English learning			−0.003***		
			(0.000)		
Starting age of English learning (<10)				−0.001***	−0.001***
				(0.000)	(0.000)
Starting age of English learning (>=10)				−0.003***	
				(0.000)	
Starting age of English learning (>=10 and >20)					−0.004***
					(0.000)
Starting age of English learning (>20)					−0.002***
					(0.000)
Num.Obs.	168 561	168 561	168 561	168 561	168 561
R2	0.200	0.253	0.253	0.254	0.254
R2 Within					
R2 Pseudo					
BIC	−459 206.0	−470 545.9	−470 545.9	−470 865.9	−470 946.5
Log.Lik.	229 657.167	235 387.270	235 387.270	235 553.288	235 599.609
Std.Errors	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust

** $p < 0.05$, *** $p < 0.01$

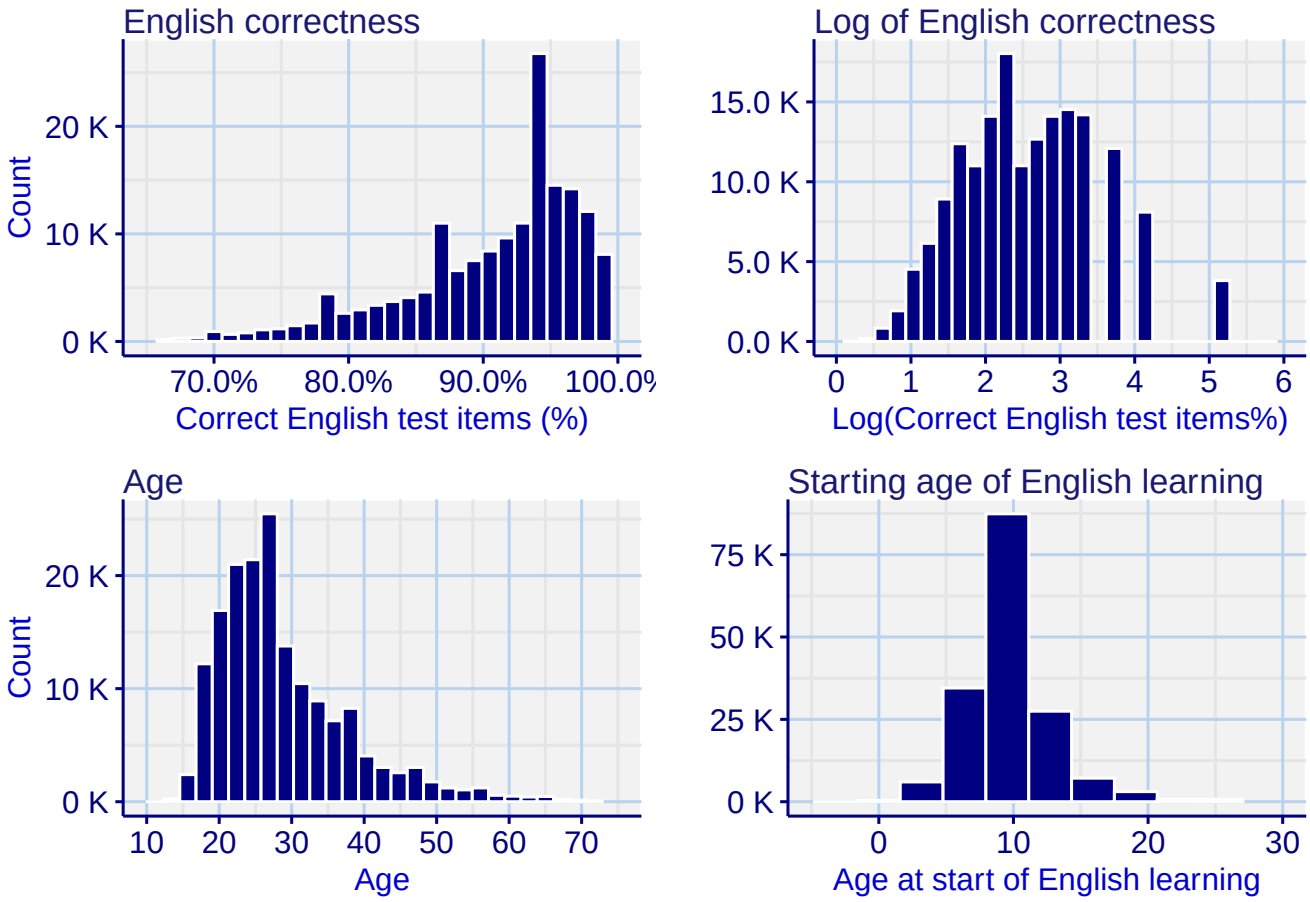

Figure 2: Distributions

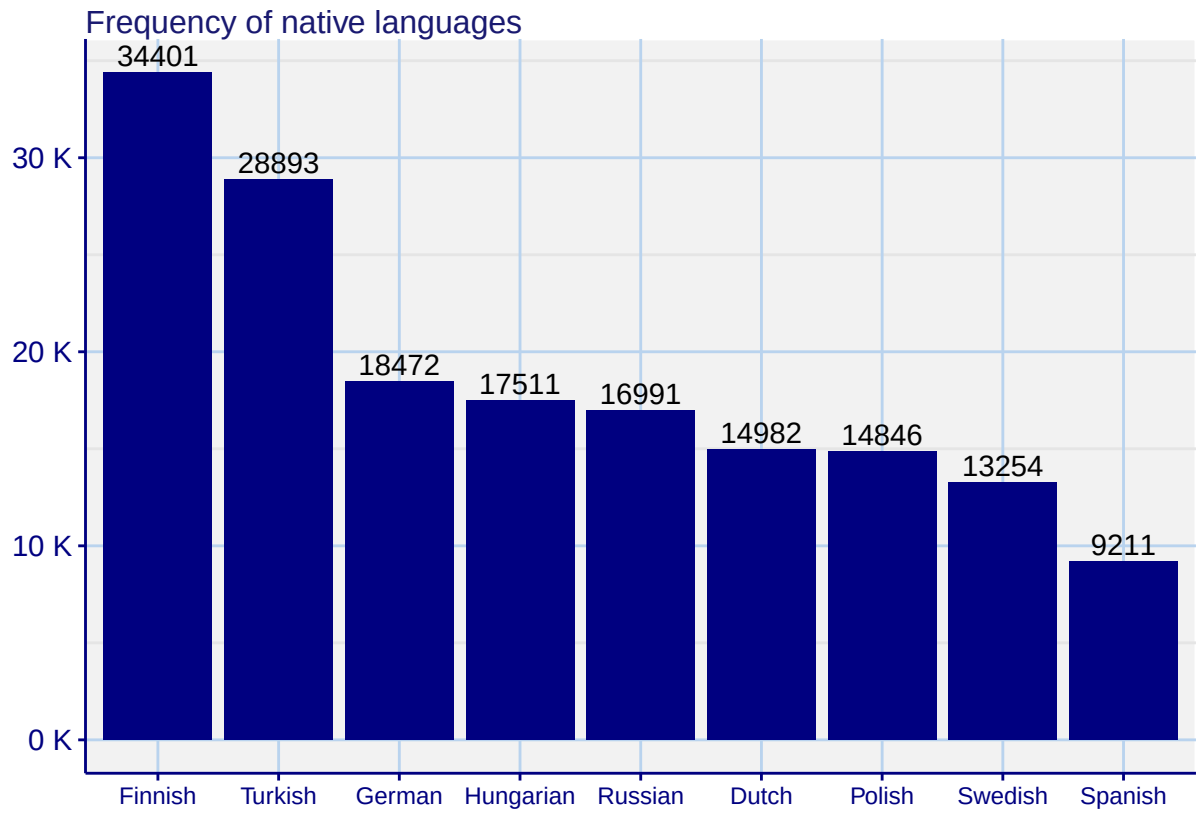

Figure 3: Frequency of native languages

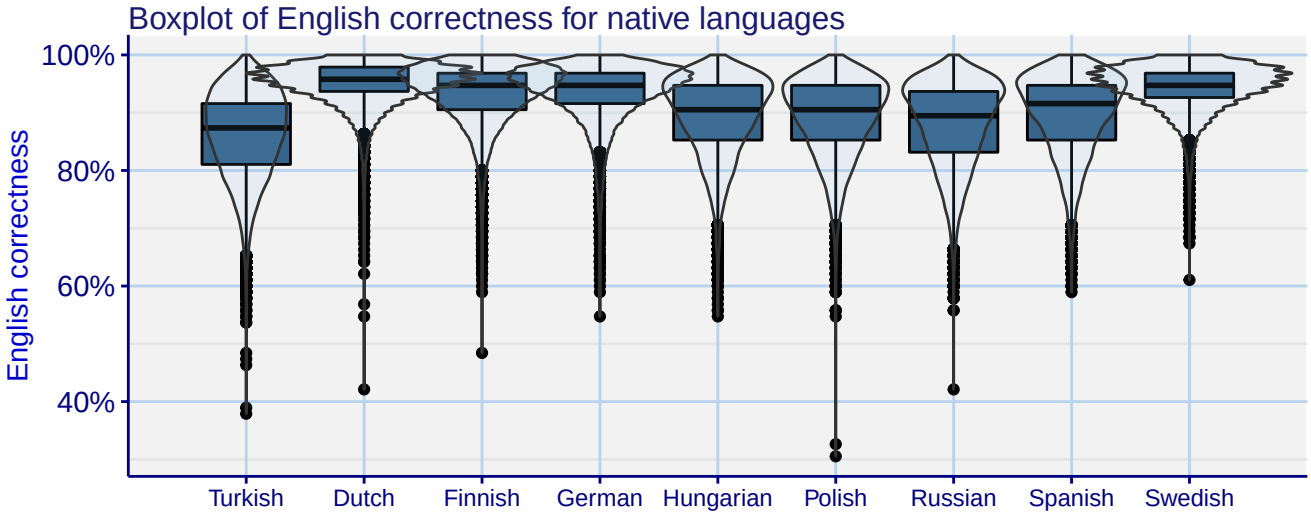

Figure 4: Boxplot of English correctness

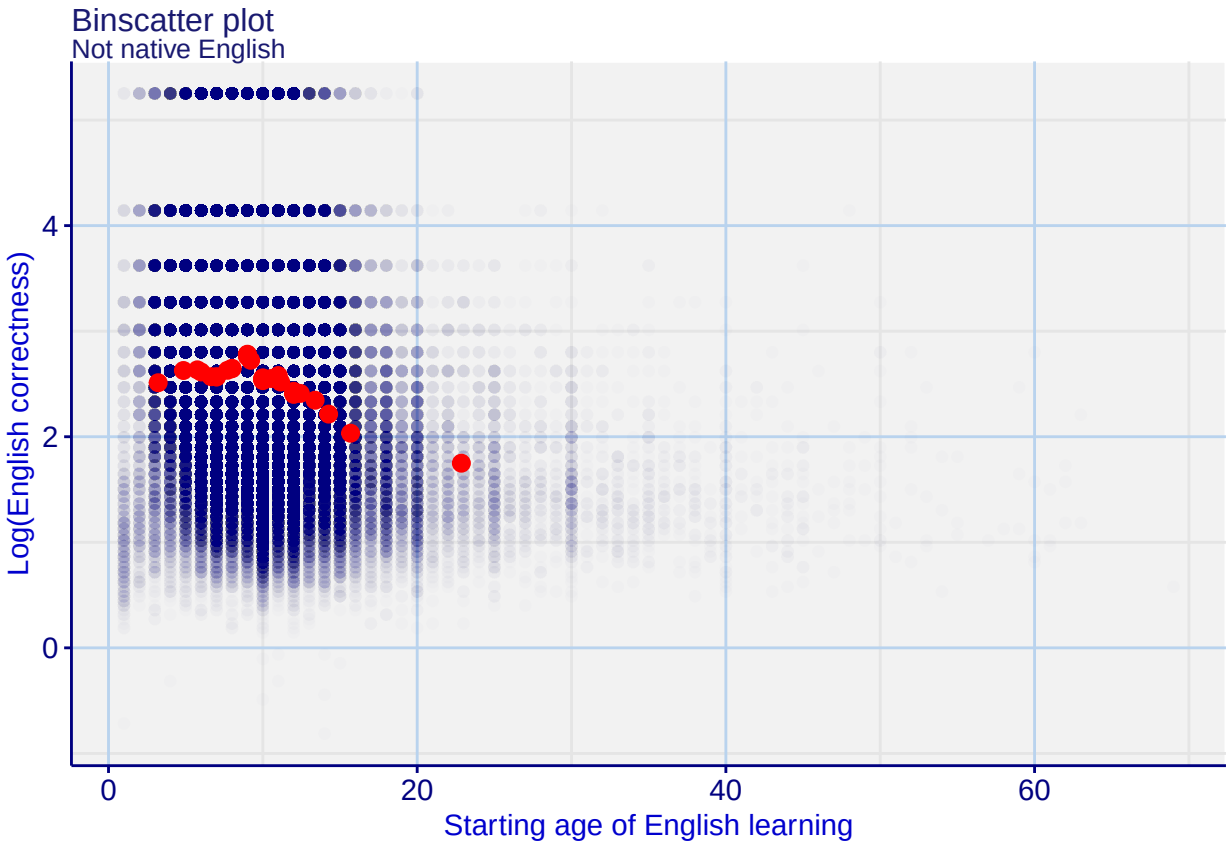

Figure 5: Binned scatterplot

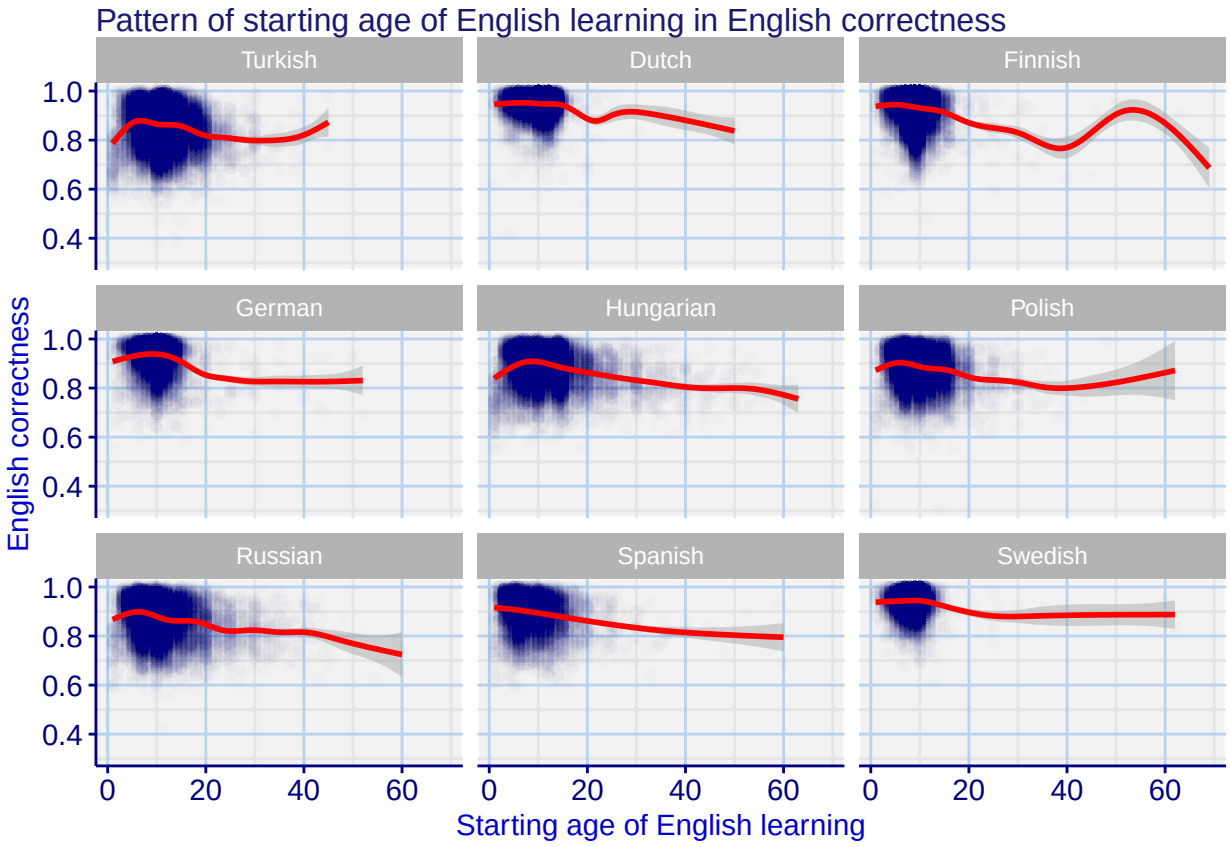

Figure 6: Smoothed pattern of age at starting of English learning

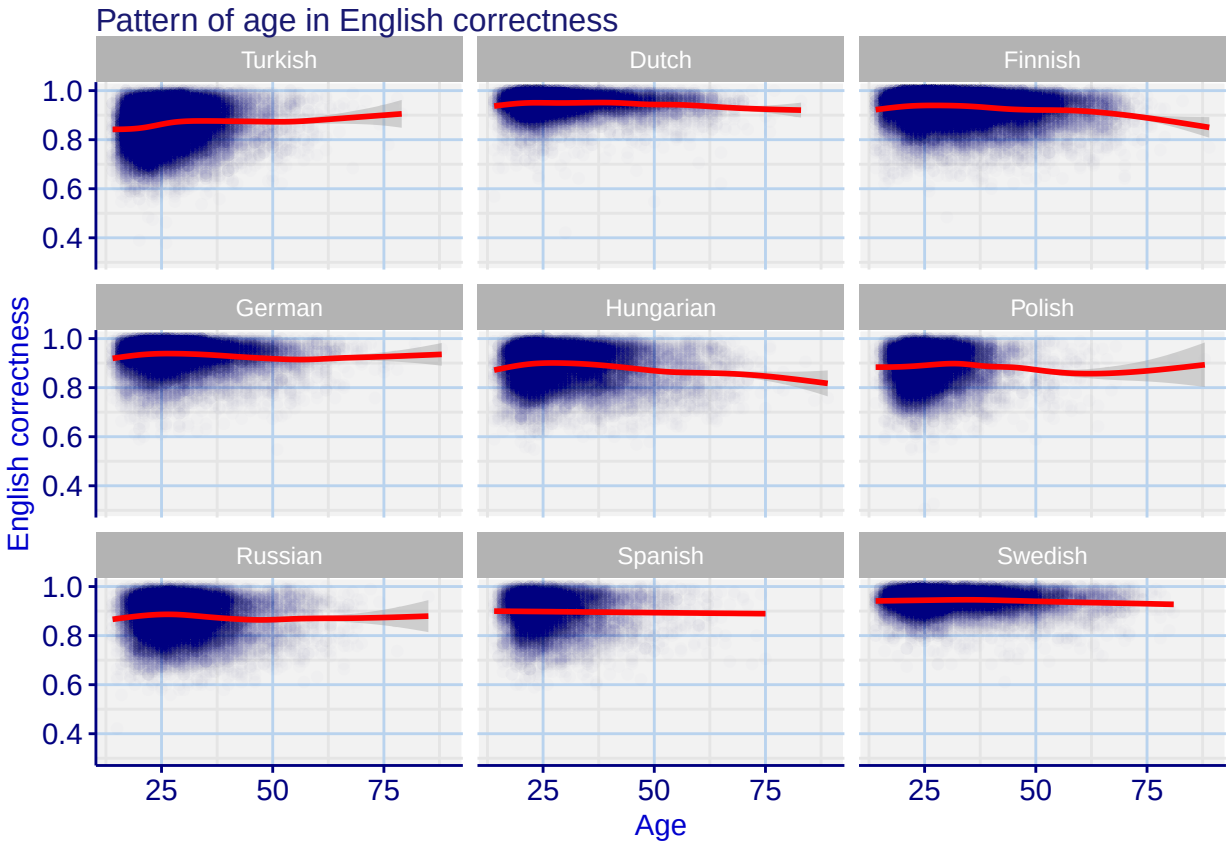

Figure 7: Smoothed pattern of age

Bibliography

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