

Acoustic Cues and Sentiment Analysis of Bilingual Speakers

Ambika Acharya, Edric Kyauk, Emily Tang

I. Introduction

Motivation

As our world becomes more international, there is an increasing interest in understanding bilingual speakers and how cultures can be seen through language. For instance, bilingual speakers have been shown to have higher cognitive control than monolingual speakers, because of their ability to switch effortlessly between two languages¹. In addition, bilingual speakers show differences in attitudes and speech when using different languages. One study found that Japanese women, when asked what they want to be in the future, answer “I want to be a housewife when I grow up” in Japanese, and “I want to be in a teacher” in English².

Acoustic features are a way to detect emotions, personality, and interpersonal stance. Previous literature reveals that automatically extracted prosodic and lexical features can be used quite accurately to predict interpersonal stance (flirtation, friendliness, awkwardness, assertiveness)³. We hope to understand bilingual speakers by analyzing acoustic features and sentiment using natural language processing techniques.

Data

The dataset we will be using is the Hamburg Adult Bilingual Language (HABLA) corpus of German-Italian and German-French bilingual speakers⁴. This dataset includes the audio (.wav files) and transcripts (.textgrid files) of 93 conversations, each 30 minutes long, between an interviewer and a bilingual speaker. Each bilingual speaker has two conversations in each language which revolve around the following subjects: place of origin and schooling, family background, food and cuisine, movies, recently read books, and cultural stereotypes spoken in German, French, and Italian.

Goal

With our motivation to understand acoustic cue and lexical differences in bilingual speakers, we want to use this data set, and see if we can accurately predict the language a bilingual speaker is using in a given conversation.

II. Methods

Our methods consist of the following: feature extraction, classification, and feature analysis. First, we extract different acoustic and textual features from the conversations sound files and transcripts. Next, we build a binary classifier using these extracted features to predict

¹ <http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00199/full>

² Ervin-Tripp, S. (1964). An analysis of the interaction of language, topic, and listener. *American Anthropologist*, 66(6_PART2), 86-102.

³ Ranganath, R., Jurafsky, D., & McFarland, D. A. (2013). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language*, 27(1), 89-115.

⁴ <http://www1.uni-hamburg.de/exmaralda/files/e11-korpus/public/index.html>

the language a bilingual speaker is using. Finally, we analyze the features using feature selection and principal components analysis (PCA), to better understand which features are important, and to uncover potential differences exhibited by bilingual speakers.

Acoustic Feature Extraction and Normalization

We extracted the following acoustic features using a PRAAT script from the .wav files: intensity (min, max, mean, sd), pitch (min, max, mean, sd), jitter, shimmer, harmonic to noise ratio, percent unvoiced, and duration. These acoustic features are commonly used in computational linguistics research.

Acoustic features vary heavily from person to person (e.g. a person may speak louder than average, a female would likely speak with a higher pitch than a male). In order to account for the different sampling population between the French-German dataset and Italian-German dataset, we normalized the acoustic features across speakers (regardless if he or she was in both datasets). We normalized these features by calculating the z-scores for each value, using the mean and standard deviation of the aggregation of the French-German and Italian-German dataset for the feature and the specific user.

Text Feature Extraction and Normalization

We extracted the following text features from the transcripts in the .textgrid files: speech rate (words/sec and characters/sec) and sentiment (proportion of positive and negative words). In order to extract the sentiment from our datasets, we used sentiment lexicons from Stony Brook University’s Data Lab⁵. We recorded how frequently a positive or negative word was used by recording the portion of positive words and the portion of negative words in a segment of conversation. Since the lexicons were of different lengths, we also divided this portion by the length of the lexicon in order to normalize these numbers. We then converted these values to z-scores so that the mean of all the values is 0.

Models: Logistic Regression and Support Vector Machines

We aggregate the data into the two bilingual speaker groups in the data set: French-German and Italian-German. With French-German, we build a binary classifier using the extracted acoustic and text features on our training data (75%) to predict whether the language of a given conversation sample is French or German (and like-wise for Italian-German). We then test our models using held-out test data (25%), as well as run 10-fold cross validation in order to ensure our models are not overfitting and can generalize.

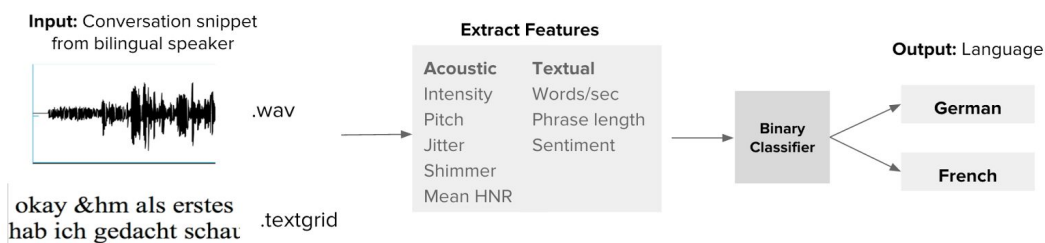


Fig. 1. Using a French-German binary classifier to predict the language

⁵ <https://sites.google.com/site/datascienceslab/projects/multilingualsegment>

We build the two models, French-German and Italian-German, using both **logistic regression** and **support vector machines (SVM)**, to see if one model can better fit our data. SVM gives us the flexibility of kernels and projecting our data into higher order dimensions, while logistic regression could produce good results if our data follows some logistic curve. We compare the two classifier accuracies in our experiments later.

Multi-class classification: German, French, Italian

In addition, we build a 3-class model to see if we can accurately predict German, French, and Italian, regardless of bilingual speakers. We hope to determine whether there are any general acoustic and textual differences across the languages, that could possibly explain any differences we find among French-German and Italian-German bilingual speakers. We build the 3-class model using both logistic regression and support vector machines, as they implement two different strategies. Logistic regression uses the “one against all” strategy, which builds a binary classifier for each class, that distinguishes one class from all other classes. SVM, on the other hand, implements the “one against one” strategy, which constructs an SVM for each pair of classes.

III. Results and Analysis

Performance of Acoustic Features

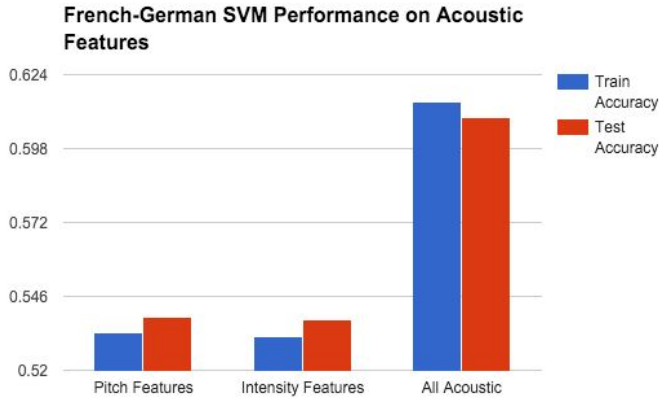
Initially, we evaluated our models when run on isolated sets of features: pitch (standard deviation, mean, max, and min), intensity (standard deviation, mean, max and min) and then with all of the acoustic features (pitch features, intensity features, jitter, shimmer, mean harmonic to noise ratio). We wanted to see if either of these subsets performed better or if including all of the acoustic features was better for the model.

French-German Accuracies for Logistic Regression and SVM on Acoustic Features

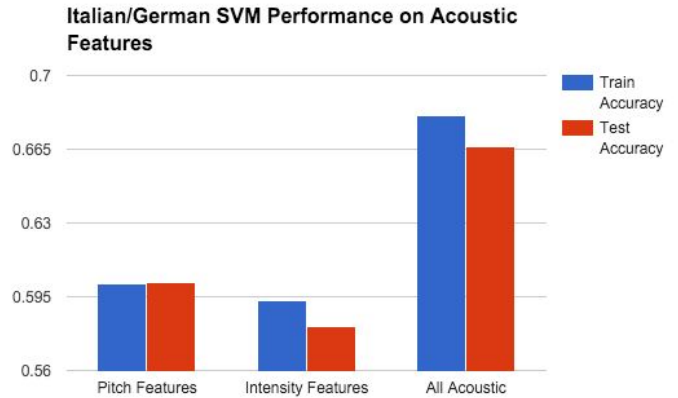
	LogReg Train	LogReg Test	SVM Train	SVM Test
Pitch Features	0.5364	0.5468	0.5333	0.5386
Intensity Features	0.5572	0.5632	0.5317	0.5379
All Acoustic	0.5811	0.5958	0.6144	0.6087

Italian-German Accuracies for Logistic Regression and SVM on Acoustic Features

	LogReg Train	LogReg Test	SVM Train	SVM Test
Pitch Features	0.6082	0.6102	0.6009	0.6017
Intensity Features	0.6034	0.5878	0.5931	0.5809
All Acoustic	0.6609	0.6563	0.6809	0.6661



Chance performance is 53.3% (predicting all German labels). The y-axis shows classification accuracy.



Chance performance is 53.9% (predicting all Italian labels). The y-axis shows classification accuracy.

We found that intensity features performed slightly better than pitch features for French/German bilingual speakers (correctly predicting the language spoken at the accuracies on the y-axis), while the opposite was true in Italian/German bilingual speakers. This suggests that pitch features might be more significant for the Italian/German model, while the intensity features might be more significant for the French/German model. However, for both datasets we found that using all of the acoustic features together gave better performance.

Performance of Textual Features

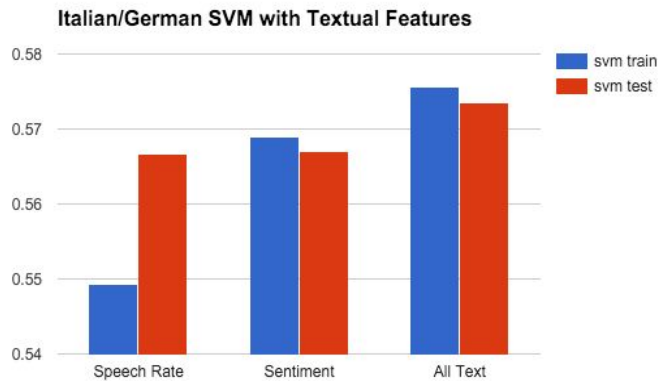
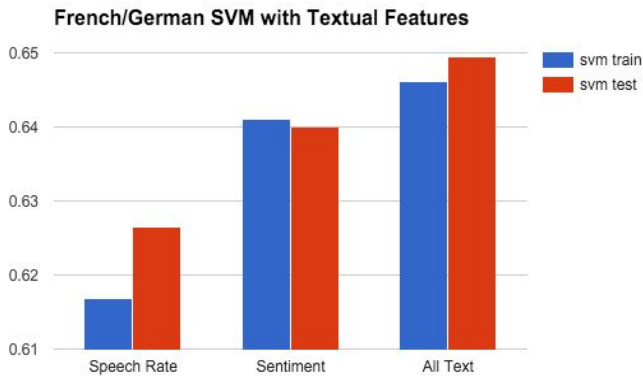
We went on to perform a similar analysis with our textual features by comparing speech rate, positive words/negative words sentiment feature and then all the textual features (the both of them) together.

French-German Accuracies for Logistic Regression and SVM on Text Features

	SVM Train	SVM Test	LogReg Train	LogReg Test
Speech Rate	0.6169	0.6264	0.6083	0.6068
Sentiment	0.6410	0.6400	0.6403	0.6331
All Text	0.6461	0.6495	0.6365	0.6400

Italian-German Accuracies for Logistic Regression and SVM on Text Features

	SVM Train	SVM Test	LogReg Train	LogReg Test
Speech Rate	0.5493	0.5666	0.5403	0.5540
Sentiment	0.5690	0.5670	0.5583	0.5493
All Text	0.5756	0.5736	0.5593	0.5632

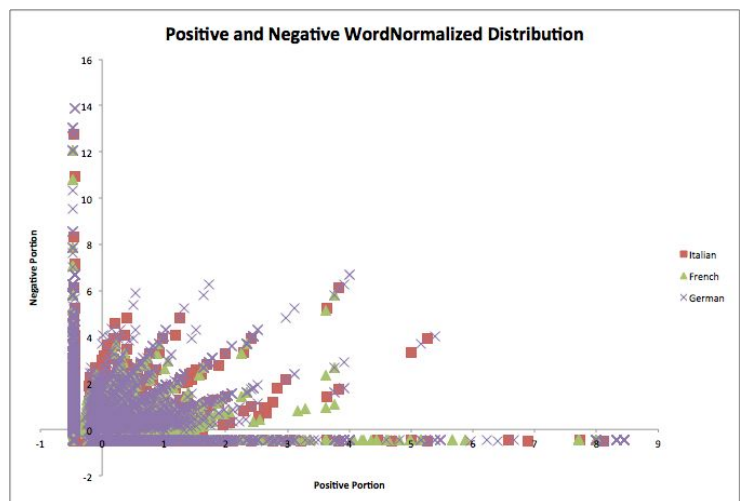


Analysis was also done using logistic regression for both datasets. The y-axis shows classification accuracy.

From this analysis we found that speech rate was more significant for French/German speakers in comparison to the effects of sentiment. However, for both languages using all of the textual features was very important for classification accuracy.

Sentiment Analysis

We found that including the portion of positive and negative words increased the accuracy. This potentially has cultural implications about the tendency to speak with a particular sentiment (or the lack of) when a bilingual speaker uses either French, German, or Italian. Looking at the graph to the right, we found that when bilingual people speak German, they tend to be more balanced with positivity and negativity than both French and Italian (the data points are clustered more along the identity line). When bilingual speakers are speaking Italian or French, they gravitate more towards using positive words than when speaking German.



Putting it all Together

After analyzing all of the features in subgroups, we ran all of the features together and compared this against our performance when running just acoustic features and just textual features. We see that having all of the features is definitely the best for our model for both datasets but find some interesting insights from looking at the breakdowns in comparison to all of the features. Acoustic features are much more significant for Italian/German bilingual speakers while textual features seem to be important for French/German speakers. This implies that in addition, sentiment features seem to help a lot more for French/German speakers since the accuracy without sentiment for French/German speakers was dramatically lower than the accuracy with all of the features. For Italian/German speakers, however, the accuracies as a

result of all features was similar to that of acoustic suggesting that acoustic features were most important for Italian/German bilingual speakers rather than textual features.

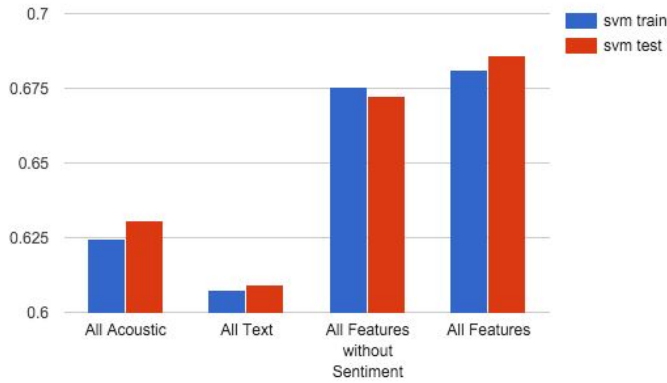
French-German Accuracies Across Acoustic and Text Features

	SVM Train	SVM Test	LogReg Train	LogReg Test
All Acoustic	0.6246	0.6308	0.6235	0.6460
All Text	0.6073	0.6090	0.6031	0.6119
All Features without Sentiment	0.6754	0.6726	0.6430	0.6555
All Features	0.6811	0.6858	0.6521	0.6602

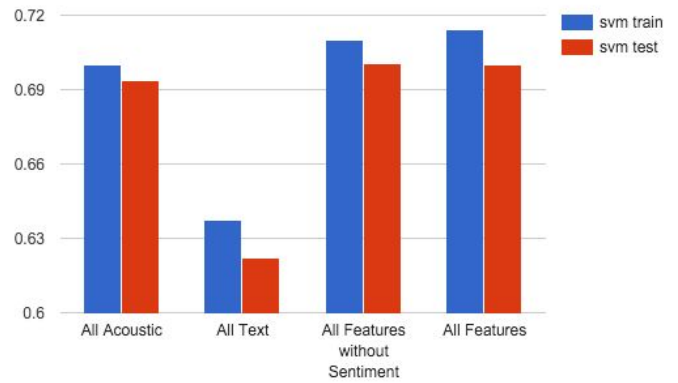
Italian-German Accuracies Across Acoustic and Text Features

	SVM Train	SVM Test	LogReg Train	LogReg Test
All Acoustic	0.6999	0.6939	0.6587	0.6494
All Text	0.6375	0.6222	0.5493	0.5363
All Features without Sentiment	0.7098	0.7006	0.6715	0.6576
All Features	0.7140	0.7003	0.6744	0.6630

French/German SVM General Analysis



Italian/German SVM General Analysis



Analysis was also done using logistic regression for both datasets. The y-axis shows classification accuracy.

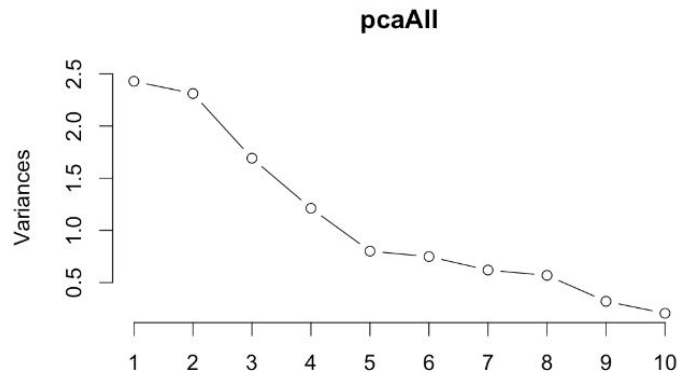
K-Fold Cross Validation

To ensure that our model did not overfit the training data and can generalize, we ran 5- and 10-fold cross validation, and report the accuracies below.

French-German SVM			Italian-German SVM		
	5-fold CV	10-fold CV		5-fold CV	10-fold CV
Train Accuracy	0.6895	0.6865	Train Accuracy	0.7204	0.7156
Test Accuracy	0.6047	0.6261	Test Accuracy	0.6376	0.6496

Principal Component Analysis

To better analyze our features, we ran principal component analysis (PCA) on a large set of our features to statistically determine which ones would be best used in our model. We ran PCA on pitch mean, pitch max, pitch min, pitch sd, intensity mean, intensity max, intensity min, intensity sd, mean harmonic to noise ratio, words per second, characters per second, duration and found our PCA distribution.



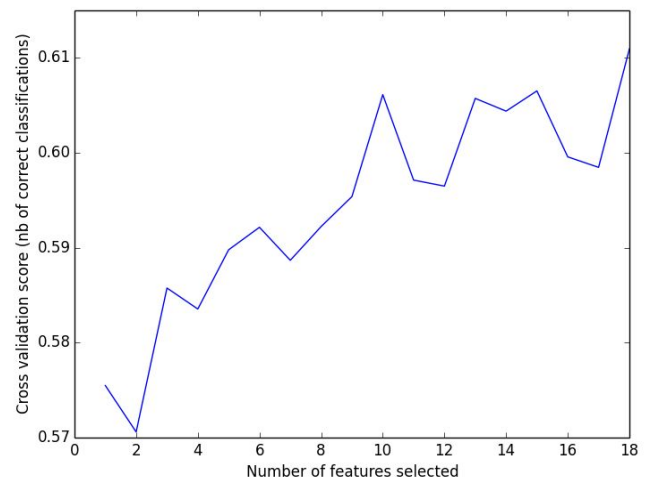
From this graph we see how much of the variance the principal components account for and see that after the fifth component the components' addition to the variance taper off since the graph plateaus. This suggests that the first five components are most important in distinguishing languages from each other. We analyze the results of PCA to determine the linguistic assumptions that correspond to these principal components (See Appendix). We evaluate the scores of PCA for the top five components and describe them as follows:

- PC1:** low intensity and level pitch
- PC2:** not very breathy (low mean_hnr), level intensity
- PC3:** slow speech rate and high average intensity
- PC4:** low pitch ceiling and low average pitch (overall low pitch)
- PC5:** level pitch, low avg intensity, and slow speech rate

Using these 5 principal components, we get the following features that we need to include in our model: imax, psd, mean_hnr, isd, words_per_sec, imean, pmax, pmean. We put these back into our logistic regression model to determine the importance of each of them on predicting language and found that all of them except intensity sd(isd) had statistically significant p-values (See Appendix).

Feature Selection

We also ran a recursive feature selection algorithm by running our model multiple times and incrementally removing the feature with the lowest weight (linear coefficients) to determine which features yielded the best accuracy with 10-fold cross validation as well. This algorithm determined that of the combination of features tried our model has the best



accuracy when using 11 features (pmin, pmax, pmean, psd, imin, imax, imean, shimmer, duration, phrase_length, neg_portion). However, since the algorithm doesn't consider all possible combinations of features, we identified that there are other features to consider in combination with these 11 features.

Considering the results of PCA and the feature selection algorithm, we found that the most important features were pos_portion, neg_portion, imin, imax, imean, isd, pmin, pmax, pmean, psd, shimmer, jitter, mean_hnr, duration, pctUnvoiced, char_per_sec, words_per_sec.

Multiclass Performance

SVM 3-Class Model Test Accuracy

	Number Predicted	Number Correct	Total Test Labels	Accuracy (Number Correct / Total)
German	2031	1458	3110	0.4688102894
French	1751	773	1485	0.5205387205
Italian	2548	1151	1735	0.6634005764

The table above shows our results from building a 3-class model using our entire data set, and all extracted acoustic and text features. We see that our model does better than chance performance because it does not predict all German labels (the class with the most number of samples). In addition, we find that our classifier is best at predicting the Italian language, with an accuracy of 66.3%. This correlates with our binary classifier performance, as our Italian-German model performs better than French-German. Overall, this implies that the Italian language is more distinct in its acoustic and textual features from German and French. This is not surprising, as it has been found that the Italian language is rich in gesture as well; Italians gesture more frequently than Americans even as children⁶.

IV. Conclusion

From our work we found that differentiating languages between bilingual speakers is a challenging task. However, we were able to find distinguishing characteristics between the languages which helped our linguistic understanding of French, Italian and German. We found that sentiment was very important in distinguishing French and Italian from German and that acoustic features were very helpful for identifying Italian conversations. We also learned a lot about using computational phonetic tools to analyze audio and transcript data, and about the complexities that can arise around feature and model selection. We hope that our work will help in understanding language differences to aid cross-cultural communication.

V. Acknowledgements

We would like to thank Rob Voigt from the NLP Group for his guidance and mentorship throughout this process.

⁶ <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2744975/>

VI. Appendix

Dataset

<http://www1.uni-hamburg.de/exmaralda/files/e11-korpus/public/index.html>

Acoustic and Text Features

<i>Acoustic Features</i>		<i>Textual Features</i>	
imin	Minimum intensity	char_per_sec	Number of characters per second
imax	Maximum intensity	words_per_sec	Number of words per second
imean	Mean intensity	Sentiment Features	
isd	Standard deviation of intensity	pos_portion	Portion of positive words in the segment
pmin	Minimum pitch	neg_portion	Portion of negative words in the segment
pmax	Maximum pitch		
pmean	Mean pitch		
psd	Standard deviation of pitch		
shimmer	Average amplitude difference divided by the average amplitude		
jitter	Local non-periodicities		
mean_hnr	Mean harmonic noise ratio (breathiness)		
duration	Duration of the segment		
pctUnvoiced	Percent unvoiced in the segment		

Results from PCA

Rotation:

	PC1	PC2	PC3	PC4	PC5	PC6
pmin	0.36544915	-0.284329816	0.16694162	-0.01072164	0.040312996	0.55111667
pmax	0.37456731	0.163160022	0.08932805	-0.51244268	-0.050124322	-0.29454035
pmean	0.37861210	-0.007776292	0.30876539	-0.43422871	-0.366280160	0.06735650
psd	-0.40456790	0.203510086	0.05449417	-0.23672479	-0.482456681	0.46226282
imin	-0.08194401	0.444791101	-0.35918389	-0.16937893	0.170415398	0.46100686
imax	-0.41314483	-0.190903985	0.35206473	0.12962817	-0.451622875	-0.14785477
imean	-0.27227775	0.153245806	0.38142560	-0.27942004	0.359312552	-0.18191795
isd	-0.15579528	-0.474576661	-0.32909481	-0.38201514	0.118286656	-0.01348021
mean_hnr	-0.22355027	-0.484636819	-0.27590790	-0.35696866	-0.003229225	-0.01314251
words_per_sec	0.29057778	-0.072821034	-0.39812995	0.24349627	-0.449904949	-0.13142968
char_per_sec	-0.09435546	0.357619978	-0.35279054	-0.19098034	-0.224521819	-0.32585563

Logistic Regression French-German Model P-Values

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.150422	0.082708	-1.819	0.068956	.
imax	0.170771	0.047705	3.580	0.000344	***
psd	-0.264501	0.036595	-7.228	4.91e-13	***
mean_hnr	0.020961	0.009048	2.317	0.020531	*
isd	-0.108514	0.027137	-3.999	6.37e-05	***
words_per_sec	0.065502	0.018719	3.499	0.000467	***
imean	0.216381	0.044396	4.874	1.09e-06	***
pmax	-0.080221	0.034759	-2.308	0.021005	*
pmean	-0.173718	0.028022	-6.199	5.67e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Logistic Regression Italian-German Model P-Values

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.150422	0.082708	-1.819	0.068956	.
imax	0.170771	0.047705	3.580	0.000344	***
psd	-0.264501	0.036595	-7.228	4.91e-13	***
mean_hnr	0.020961	0.009048	2.317	0.020531	*
isd	-0.108514	0.027137	-3.999	6.37e-05	***
words_per_sec	0.065502	0.018719	3.499	0.000467	***
imean	0.216381	0.044396	4.874	1.09e-06	***
pmax	-0.080221	0.034759	-2.308	0.021005	*
pmean	-0.173718	0.028022	-6.199	5.67e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1