Research Project

|  |
| --- |
| Francisco Orejarena1 & Anya Lunden2 |
| 1 Rutgers University |
| 2 College of William and Mary |
|  |

# Author note

Here is my author note(?): I go to Rutgers.

I don’t know what an author note, and frankly, I’m too nervous to ask.

Correspondence concerning this article should be addressed to Francisco Orejarena, New Brunswick, New Jersey. E-mail: [batfran0621@gmail.com](mailto:batfran0621@gmail.com)

Abstract

Before I write the abstract, I don’t really know what indentation that the Markdown file keeps mentioning; I just felt that was worth mentioning. Anyway, this paper is an analytical discussion of the Functional Load Hypothesis (FLH). Using the Stress Correlate database, linguist Anya Lunden sought to disprove the FLH. This paper will expand on the analysis conducted by Lunden.

*Keywords:* phonetic correlates, stress, Functional Load Hypothesis

Word count:

Research Project

There are a couple of things that I would like preface: the first is that since this is not my data, most of my information will come from the analysis, results, and discussion section. This is important, as that is where the conclusion on FLH comes from. The second aspect of this paper is that I am unsure of how to format this paper without messing up the code. Regardless, enjoy the paper!

# Methods

As Anya Lunden so aptly put it, “The functional load hypothesis of Berinstein (1979) put forward the idea that languages which use a suprasegmental property (duration, F0) contrastively will not use it to realise stress.” (2017). But, with a wealth of linguistic knowledge more recordable than evr before, does the theory hold up? Lunden did not think so in her paper “Vowel-length contrasts and phonetic cues to stress: an investigation of their relation” (2017), and I am inclined to agree. After reading Lunden’s paper, while her analysis and data was not done badly per say, the comparative models could have been done more easily. I will strive to do so, and support Lunden’s conclusion that the FLH is not a universal standard, and is a presumptuous theory at best.

The methods used in this paper were statistical anylases; these were some nested model comparisons, followed by anova analyses. The Stress Correlate Database was used as the data points in the analysis. The materials are the sorted version of the Stress Correlate Database, originally put together by Lunden, and then revised by myself.

## Participants

There were no participants in this paper. However, there were 140 data points (languages) used in the analyses. These will be discussed in the materials and procedural section.

## Material

The material used in this analysis is from the Stress Correlate Database. First, the data, was sorted by Lunden herself. Lunden’s data was sorted into an excel spread sheet, with the following criteria:

1. Language information

* Language name
* Primary family
* Secondary family
* ISO 639-3 code

1. Stress system information

* Stress system type (single stress, binary stress, dual stress, ternary stress)
* Primary stress edge (left/right)

1. Stress correlate information

* Duration as reported correlate (yes/no)
* Pitch as reported correlate (yes/no)
* Intensity/loudness/energy/spectral tilt as reported correlate (yes/no)
* Primary correlate, if known
* Vowel reduction (phonological) in unstressed syllables (yes/no)

1. Phonological uses of suprasegmentals

* Duration used contrastively outside of stress (yes/no)
* Pitch used contrastively outside of stress (yes/no)

1. References: Sources consulted (full bibliographical entries are given below)

This list of information and the sorted and organized spreadsheet was a little complicated and redundant. My reorganized list is as follows:

1. Language information

* Language name
* Primary family

1. Stress system information

* Stress system type (single stress, binary stress, dual stress, ternary stress)

1. Stress correlate information

* Duration as reported correlate (yes/no)
* Pitch as reported correlate (yes/no)
* Intensity/loudness/energy/spectral tilt as reported correlate (yes/no)

1. Phonological uses of suprasegmentals

* Duration used contrastively outside of stress (yes/no)

The first thing that I got rid of is section ‘e’ becuase having the references is pointless for tidying data. The next was to sort the list from top to bottom. Starting with section ‘a’, I got rid of secondary language information, as well as the ISO code. This was because secondary langauge information is irrelevant in discussing supersets, especially in terms of typological comparisons. ISO code is unecessary as it is just an ID number, essentially. Moving on to section ‘b’, I got rid of where primary stress is found, as the location is irrelevant in comparitively stress analysis; the main comparisons are centered more around the realizations more than anything else. In section ‘c’, I got rid of if the primary correlate was known, as the column was extremely sparse, with 50% of the data not being filled. This made it unreliable to use in any anylsis. I also removed vowel reduction; in my presentation on this, someone noted that vowel reduction can be a process of stress, as with the famous [shwa]. This means that some unstressed words use [shwa] as a contrastive feature to describe unstressed segments. After some careful reconsideration, I decided to keep it removed from the database parameters, as I felt it was more of an indicator of secondary stress, and would not help in proving or disproving my hypothesis. Section ‘d’ was a little sparse as well, having only two parameters for length and pitch, but with reason: well, suprasegmental features are prosodic, so this is considered to be the contrastive features that Lunden added in her parameters. However, I only kept duration, as that seemed to be the main feature discussed in the original paper by Lunden. The last things that I did was restructure the criteria and numbering of stress systems, excluding the literal column named ‘stress’. The 0-3 system was confusing, and borderline redundant: “In this paper, a ‘0’ or ‘3’ language is considered to lack contrastive vowel length, and a ‘1’ or ‘2’ language is considered to have contrastive vowel length.”(2017). For this reason, I simply went through, and made any ‘3’s ’0’, and any ’2’s into ’1’s. I did keep the stress columns 0-3 marking as it helped with the coloring of everything (see graph). Finally, I numbered the languages, so that they would be able to be sorted through my code. This may seem bizarre, but with the language names being present, it proved to be problematic.

## Data Analysis/Procedure

The first thing I did was import my newly organized data

## # A tibble: 6 x 9  
## X1 `Language name` `Language famil~ `Stress system ~ Duration Pitch  
## <int> <chr> <chr> <int> <int> <int>  
## 1 1 Aleut "Eskimo\x96Aleu~ 0 1 0  
## 2 2 Anejom Austronesian 1 0 1  
## 3 3 "Apuri\xf1a" Arawakan 1 0 1  
## 4 4 Arabela Zaparoan 0 1 0  
## 5 5 Arabic (Ammani-J~ Semitic 0 1 1  
## 6 6 Arabic (Syrian) Semitic 0 0 1  
## # ... with 3 more variables: Intensity <int>, `Duration =  
## # contrastive` <int>, `Pitch = contrastive` <int>

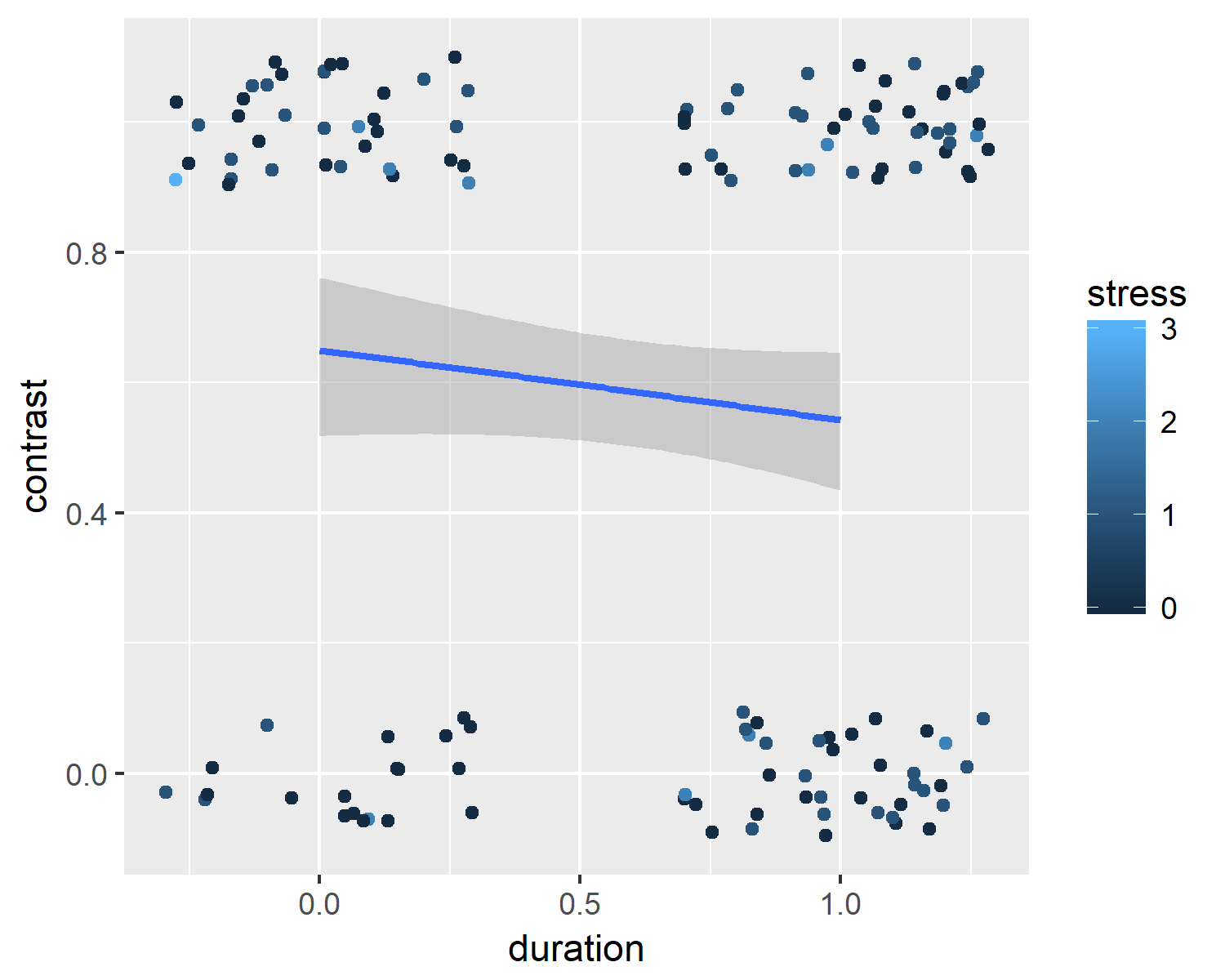
Followed by some simple tidying, which resulted in ‘lang\_all’:

## # A tibble: 141 x 6  
## lang stress duration pitch intensity contrast  
## <int> <int> <int> <int> <int> <int>  
## 1 1 0 1 0 0 1  
## 2 2 1 0 1 1 1  
## 3 3 1 0 1 0 1  
## 4 4 0 1 0 1 1  
## 5 5 0 1 1 1 1  
## 6 6 0 0 1 0 1  
## 7 7 1 1 1 1 1  
## 8 8 1 0 1 1 0  
## 9 9 2 0 1 1 0  
## 10 10 0 0 1 1 1  
## # ... with 131 more rows

From this, I plotted the data.

## Warning: Removed 1 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).



This plot shows that there is a negative correlation between duration and whether or not it is contrastive: it seems that it is variable, and more often than not, you can have duration be a contrastive feature, or you can have stress not be a contrastive feature. So far, the FLH isn’t holding up so well.

I then made models, using the general linear model, of three different variables: duration, intensity, and all varibables in the tidied data. The duration (the main data point for this comparison) was used against a null model.

##   
## Call:  
## glm(formula = contrast ~ 1, family = binomial(link = "logit"),   
## data = lang\_all)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.328 -1.328 1.034 1.034 1.034   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.3463 0.1716 2.018 0.0436 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 189.95 on 139 degrees of freedom  
## Residual deviance: 189.95 on 139 degrees of freedom  
## (1 observation deleted due to missingness)  
## AIC: 191.95  
##   
## Number of Fisher Scoring iterations: 4

##   
## Call:  
## glm(formula = contrast ~ duration, family = binomial(link = "logit"),   
## data = lang\_all)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4473 -1.2500 0.9297 1.1065 1.1065   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.6152 0.2775 2.217 0.0267 \*  
## duration -0.4461 0.3544 -1.259 0.2081   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 189.95 on 139 degrees of freedom  
## Residual deviance: 188.34 on 138 degrees of freedom  
## (1 observation deleted due to missingness)  
## AIC: 192.34  
##   
## Number of Fisher Scoring iterations: 4

##   
## Call:  
## glm(formula = contrast ~ intensity, family = binomial(link = "logit"),   
## data = lang\_all)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.341 -1.334 1.022 1.022 1.049   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.31015 0.28069 1.105 0.269  
## intensity 0.06632 0.35632 0.186 0.852  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 187.11 on 137 degrees of freedom  
## Residual deviance: 187.08 on 136 degrees of freedom  
## (3 observations deleted due to missingness)  
## AIC: 191.08  
##   
## Number of Fisher Scoring iterations: 4

##   
## Call:  
## glm(formula = contrast ~ duration + intensity + pitch, family = binomial(link = "logit"),   
## data = lang\_all)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5173 -1.2942 0.8895 1.0649 1.2550   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.31979 0.50703 0.631 0.528  
## duration -0.40487 0.38145 -1.061 0.289  
## intensity -0.09557 0.37359 -0.256 0.798  
## pitch 0.45119 0.38529 1.171 0.242  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 184.96 on 135 degrees of freedom  
## Residual deviance: 181.63 on 132 degrees of freedom  
## (5 observations deleted due to missingness)  
## AIC: 189.63  
##   
## Number of Fisher Scoring iterations: 4

## Results

Looking at the summaries, the only significant model seems to be the general linear model of duration, as the p value is smaller than 0.05. With the intercept a little tricky to interpret, I would say it most likely represents a language with none of the variables as a correlate, and would mean that there is a significance of duration and an abnormal occurence (relative to the other correlates) of duration in the world’s languages. This would probably mean that the larger p-values disprove the FLH, as there is no significance in them.

# Discussion

The Fucntional Load Hypothesis was a popular theory early on in phonology. It made logical sense; how could you have a language utilize multiple correlates of stress? Well, the more we have learned about language, the more that people have realized that it doesn’t matter how complex it may seem, it won’t stop a language from being so complex (ergative languages, anybody?). The analysis of the nested model comparisons is probably the most important part of this: while Lunden did a good job, and her work is most definitely valuable, she could’ve made easier models to compare the database. I would also like to acknowledge, as an extension of that, while my desire to work on this database was a result of this necessity of model comparisons, it was extremely difficult, and showed why Lunden approached it the way that she did. The work on these 140 languages was thorough, and showed how complex languages are, as well as their suprasegmental features. Even in a simple overview of the database, one can see that it is more than an abnormality or statistical outliers that the languages do not coincide with FLH, and that maybe more theories need to be reevaluated in the future.

# References

Lunden, A., Campbell, J., Hutchens, M., & Kalivoda, N. (2017). Vowel-length contrasts and phonetic cues to stress: an investigation of their relation. Phonology, 34(3), 565-580.