A Bootstrap-Based Reanalysis of Zamma (2013)

Shigeto Kawahara

Keio University

The Institute of Cultural and Linguistic Studies

Abstract

In his recent book "Patterns and Categories in English Suffixation and Stress Placement: A Theoretical and Quantitative Study," Zamma (2013) identified four classes of English suffixes in terms of the (i) root attachment behaviors and (ii) stress patterns, instead of the more traditional "Class 1 vs. Class 2" distinction (Siegel, 1974). He showed that these four types of suffixes are not evenly distributed in the English lexicon, and their distributions are affected by whether the suffixes are light or heavy. He went on to argue that with the theory of unranked constraints developed by Anttila (2002), we can predict these distributions. This short paper offers a statistical reassessment of this claim using a bootstrap resampling method.

Background

Hideki Zamma passed away on March 22nd, 2016, a week after he was involved in a car accident, at the age of 46.[1] In his recent book "Patterns and Categories in English Suffixation and Stress Placement: A Theoretical and Quantitative Study" (Zamma, 2013), he identified four classes of English suffixes in terms of the (i) root attachment behaviors and (ii) stress patterns, instead of the more traditional "Class 1 vs. Class 2" distinction (Siegel, 1974). He showed that these four types of suffixes are not evenly distributed in the English lexicon, and their distributions are affected by

[1]The obituary is provided in Appendix A.

whether the suffixes are light or heavy (taking into consideration the "extrametricality" and other effects). On page 131 of the book, he summarizes the distributions of these suffixes in the English lexicon.

Table 1: The distributions of the four types of suffixes identified in Zamma (2013)

	Class 1	Class 2	Class 3	Class 4
L	24	16	18	0
H	37	3	18	3

This is the core descriptive finding of his book. He went on to argue that with the theory of unranked constraints developed by Anttila (2002), we can predict the distributions in Table 1. Given his assumptions about the constraint set and which constraints are unranked with respect to each other, he argued that the predicted distributions should be as follows (p. 133):

Table 2: The predicted distributions

	Class 1	Class 2	Class 3	Class 4
L	20.7	10.8	19.6	6.9
H	28.4	5.9	19.2	7.5

He used a Chi-square test to compare Tables 1 and 2 and to examine the goodness of fit between them. This approach has a few problems: (i) we would have to rely on null results, (ii) we cannot figure out which classes fit the prediction and which classes do not, and (iii) we do not know the properties of the distribution underlying this sample. I suggested to him in the summer of 2015 that he should instead run a bootstrap resampling test.

A boostrap reanalysis

Now he has left us, before implementing this bootstrap analysis together, so I would like to do it here as a belated tribute to him. Based on the data in Table 1, I created 50,000 new samples using resampling with replacement (Efron & Tibshirani, 1993). We can calculate 95% percentiles over these samples to get 95% confidence intervals. I used R to implement this bootstrap method: the code is provided in Appendix B.

Table 3: The 95% confidence intervals calculated by the bootstrap method

	Class 1	Class 2	Class 3	Class 4
L	17-31	10-23	11-25	0
H	29-44	0-7	11-25	0-7

Class 4 in the L category does not exist in the English lexicon, so let us leave that aside.[2] The question is whether each value in Table 2 fits within the 95% confidence intervals calculated in Table 3. And they generally do. The H-suffixes for Class 1 and Class 4 are a bit off, but the rest fit in the predicted ranges, and even those that do not fit are very close.

I am not an expert in this area, but am glad to find that his results are as good as this. I would like to dedicate this result to his memory.

References

Anttila, Arto (2002) Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory* **20**(1): 1–42.

Efron, Bradley & R. J. Tibshirani (1993) *An Introduction to Bootstrapping*. Boca Raton: Chapman and Hall/CRC.

Siegel, Dorothy (1974) *Topics in English Morphology*. Doctoral dissertation, MIT.

Zamma, Hideki (2013) *Patterns and Categories in English Suffixation and Stress Placement: A Theoretical and Quantitative Study*. Tokyo: Kaitakusha.

Appendix A

The following is the obituary that I posted on phonolist, an online blog for phonologists (`http://bit.ly/1NSCwUl`). People are welcome to contribute their memories with him using the comment function.

[2] One could of course argue that if the absence of the L Class 4 suffix is a true grammatical gap, the theory should not allow its presence at all, unlike the one presented in Table 2.

Hideki Zamma passed away on March 22nd, 2016, a week after he was involved in a car accident, at the age of 46. Hideki was a very active phonologist, and professor at Kobe City University of Foreign Studies. His research focused on phonological variation and formal phonological theory. He worked on various topics including rendaku, Japanese accent, English stress, and formal properties of local conjunction. His recent book "Patterns and Categories in English Suffixation and Stress Placement: A Theoretical and Quantitative Study" (Kaitakusha, 2013), based on his PhD thesis submitted to Tsukuba University (2012), explored item-specific behaviors of different English suffixes within the framework of unranked constraints in Optimality Theory, which won a prize from the English Linguistic Society of Japan as well as Ichikawa Prize. In addition to his research, he served as a board/organizing/editorial member for the Phonological Society of Japan, the Phonetic Society of Japanese, and the English Linguistic Society of Japan. In addition to being a great researcher, he was also a caring and dedicated teacher. Hideki kept trying to make linguistic materials as assessable as possible, for example, by teaching distinctive features based on "slips of the ear" patterns, using a famous Japanese TV show. He will be greatly missed by his family, friends, colleagues, and students.

Appendix B: The R code

```
# This program uses a Bootstrap method and calculates 95% percentile intervals of all the elements in a sample. The
 number of the original sample is n; resampling is repeated m times.

class=c("c1","c2","c3","c4") # define all the elements
probabilities=c(0.606557377,0.049180328,0.295081967,0.049180328) # and their probabilities

n<-61 # the size of the orignal sample
m<-50000 # the number of resample

    my.boot.c1<-numeric(0)  # create a hash
    my.boot.c2<-numeric(0)
    my.boot.c3<-numeric(0)
    my.boot.c4<-numeric(0)

for(i in 1:m) {
    y<-sample(class,n,replace=TRUE,prob=probabilities) #resampling
    my.boot.c1[i]<-length(grep("c1",y)) # count the number of elements, and put it in the hash
    my.boot.c2[i]<-length(grep("c2",y))
    my.boot.c3[i]<-length(grep("c3",y))
    my.boot.c4[i]<-length(grep("c4",y))
        }

Y<-matrix(nrow=4,ncol=3) # prepare a matrix to store results

Y[1,1] <- mean(my.boot.c1) # encoding means in the first column
Y[2,1] <- mean(my.boot.c2)
Y[3,1] <- mean(my.boot.c3)
Y[4,1] <- mean(my.boot.c4)


Y[1,2-3]<-quantile(my.boot.c1,p=c(0.025,0.975)) # calculate the percentile intervals
Y[2,2-3]<-quantile(my.boot.c2,p=c(0.025,0.975)) # and store them in Y
Y[3,2-3]<-quantile(my.boot.c3,p=c(0.025,0.975))
Y[4,2-3]<-quantile(my.boot.c4,p=c(0.025,0.975))
Y # print

# If you want each value

mean<-Y[,1] # means
lower<-Y[,2] # lower bound
upper<-Y[,3] # upperbound
```