Bethany Lycan
LING 596
Final Paper

Cluster Analysis of Vowel Length in Pahka'anil

Introduction

Pahka'anil, also known by its exonym Tübatulabal (TUB, ISO 639-3), is a Uto-Aztecan language spoken by the Pakanapul tribe in the southern region of the eastern Sierra Nevada mountain range. While there are no longer any native speakers of Pahka'anil, the language is simultaneously being learned and taught by Pakanapul tribal leaders in Bakersfield and Lake Isabella.

Charles F. Voegelin produced the first major set of work on Pahka'anil when he conducted fieldwork with Tübatulabal speakers in the 1930s for his dissertation, *Tübatulabal Grammar* (1935a), accompanied by *Tübatulabal Texts* (1935b) and a *Working Dictionary of Tübatulabal* (1958). The majority of the analyses done on Pahka'anil since then have been based on Voegelin's data and the analysis given in his grammar. In the 1950s, Sydney M. Lamb and Hansjakob Seiler also conducted field work with members of the tribe, which Lamb references in his discussion of the language families in the Great Basin (Lamb 1958).

Linguist Lindsay Marean has been working with the tribe to develop pedagogical materials for the last several years. The Pakanapul Language Team (consisting of Marean and the tribal leaders involved in the revitalization project) worked with Jim Andreas, the last native speaker, to further document the language until his death in 2008 (Robert Gomez, Tribal Chair, p.c.). Among the materials developed by Marean is a *Pahka'anil-English Dictionary* (2015), in which she has compiled from a variety of sources the part of speech, English translation, examples, and all known variations for each entry.

Literature Review

In the body of work that has been done on Pahka'anil, there has been an *a priori* assumption that Pahka'anil has a distinction between long and short vowels. This assumption has been based on Voegelin's grammar in which he states that these categories exist. Throughout Marean's dictionary, however, a single entry may have several different spellings listed depending on the source.[1] For example, the entry for *induugal* 'that one' has the variants *unduugal*, *undugal*, *wündagal*, *undagal*, and *anduugal*. Likewise, *tangalangil* 'thunder' also can be written as *dawaagalanggil*, *tangalaangil*, and *taangalaangil*. These variations depend on the source of the data; each researcher transcribed the word based on what they heard, which, as seen in these examples, has led to discrepancies in the data.

Though researchers have differed in how they transcribe vowel length within the same words in Pahka'anil, no one has questioned the existence of these two categories. This pilot study investigates that assumption; it utilizes cluster analysis to test for the existence of separate long and short vowel categories.

Why is questioning this assumption important? As second language learners of their heritage language, the members of the tribe are required to learn sounds and sequences that are either not present or salient in their native language, English. For example, they have to learn how to pronounce the /ɨ/ sound and to be more conscientious of the glottal stop. English does not have a distinction between long and short vowels, and as such, Pahka'anil learners have to learn to differentiate between them. If this category does not actually exist, this is one less thing the tribe members need to worry about when trying to reacquire their heritage language.

Research Question: Does Pahka'anil have two categories of vowel duration?

[1] The orthography that has been developed by the Pakanapul Language Team distinguishes between short and long vowels by writing a single letter for a short vowel and two letters for a long vowel.

Cluster analysis is a statistical technique used to classify objects into groups with homogeneous characteristics. This is used in a variety of fields including biology, astronomy, psychology, and marketing (Everitt et al., 2001). A form of cluster analysis, though sometimes not named as such, is commonly used in descriptive linguistics when identifying the categories that exist in a language. For instance, Ahland (2009) describes the set of vowels that are present in Northern Mao by measuring formant values and lengths of each vowel, then comparing the ranges and means of the F1, F2, and length values. The vowel inventory and distinction between short and long vowels is identified through the clumps that emerge from these measurements, as well as noting the differences in the means. Similarly, Fulop & Warren (2014) use this technique in their discussion regarding the presence of advanced tongue root in vowel harmony in Karajá. They took measurements of F1, F2, and F3 values, and based their subsequent analysis on the clusters that appeared in the measurements.

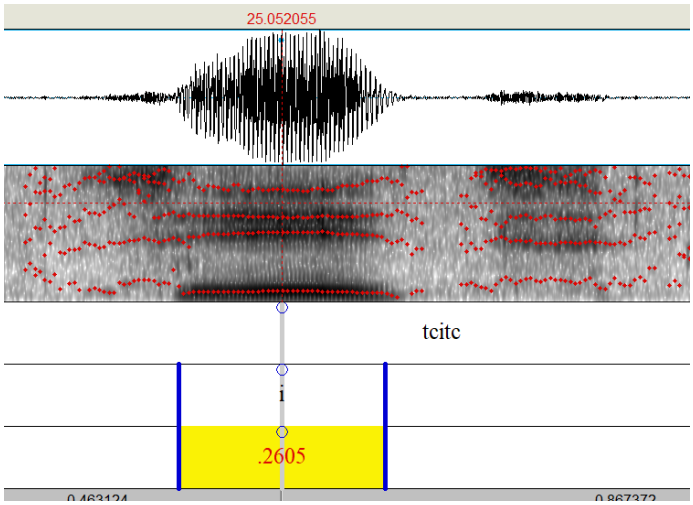
Pagano et al. (2015) explicitly use cluster analysis in their comparison of source and translated texts; they use it to compare the length and grammatical structures in the source text and the translations. Jones et al. (2012) use a k-means cluster analysis as part of their model of infants' acquisition of vowels in Gurindji Kriol. Moisl (2015) advocates an increase in the use of cluster analysis in corpus linguistics, arguing that it would improve objectivity and replicability of studies as well as aid in the discovery of patterns in large bodies of data.

Methodology

This study utilizes the recordings of the elicitation sessions conducted by Sydney M. Lamb and Hansjakob Seiler in 1954. These recordings are housed in the California Language Archive managed by the Survey of California and Other Indian Languages. To the best of my knowledge, no one has performed an acoustic analysis of this data. In particular, no one has

taken measurements of vowel duration in Pahka'anil; all transcriptions of words in the language have solely been based on what was heard by the researcher.

Forty-nine vowel duration measurements were taken from two of the Lamb and Seiler recordings (Elicitation of numbers, LA 80.008; Elicitation of words related to animals, LA 80.012). The measurements were conducted in Praat (Boersma & Weenink, 2017) following the steps outlined by Wright & Nichols (2015). Measurements of the first and second formants for each vowel were taken from the approximate midpoint of the vowel. Although this is not the focus of the present study, there is also variation in which vowel is written, as can be seen in the examples given above. Using the formant values rather than relying on the orthography will increase the validity of the classification of the vowels (or at least demonstrate the need for further acoustic analysis). The Pahka'anil words and their F1, F2, and duration measurements are listed in the Appendix (the vowels measured are in bold), and an example of one such measurement is shown in the figure below. Hierarchical clustering of the data was then conducted in SPSS.



Findings

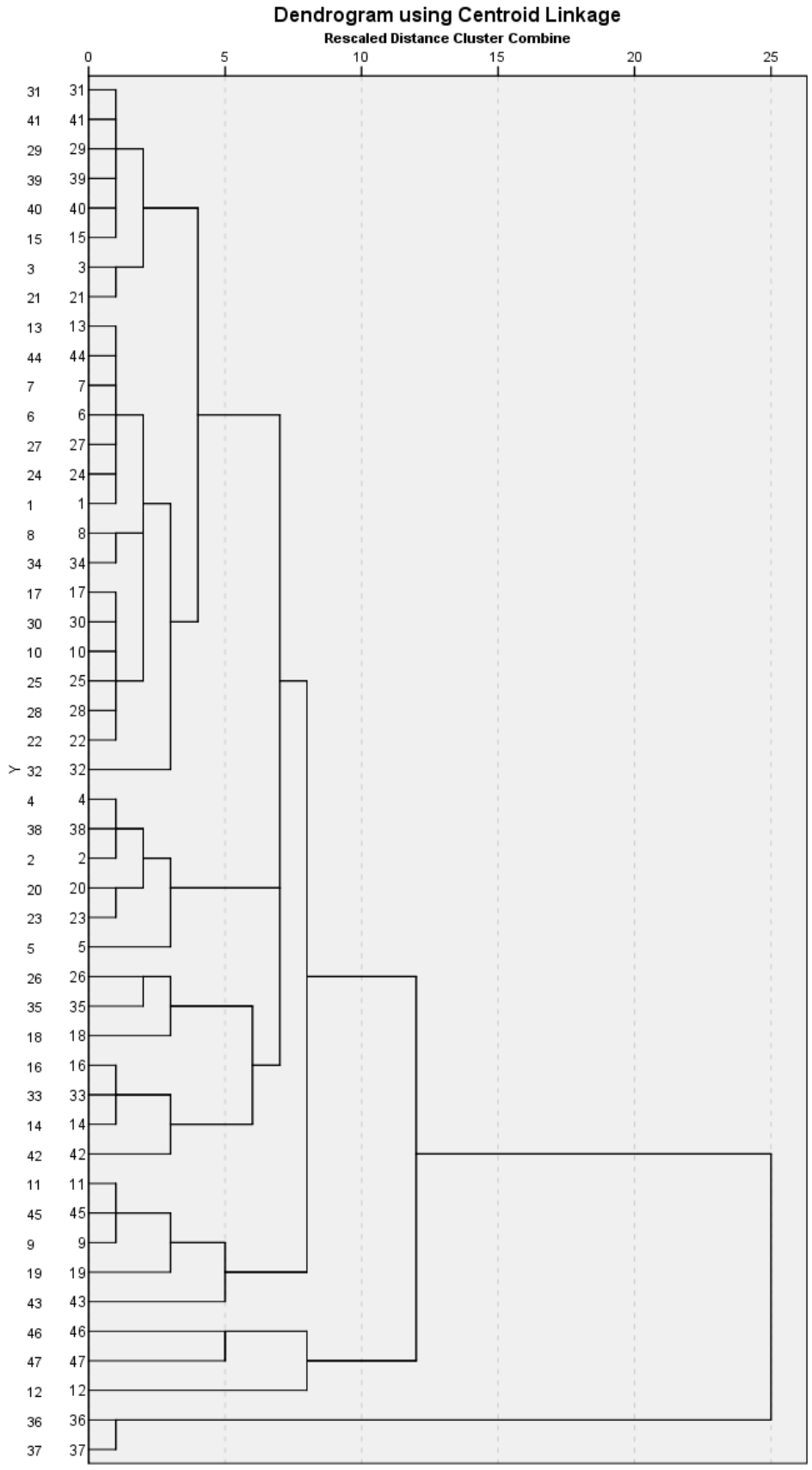
Agglomerative, hierarchical clustering was performed in SPSS based on the steps outlined in Hair et al. (1995). The clustering method selected was centroid clustering, and the distance measure used was squared Euclidean. This method was selected as it is less sensitive to potential outliers. As the variables were of different scales (the formants were in the thousands whereas the lengths were in thousandths), the values were standardized to z-scores, allowing the measurements to be equally weighted when determining clusters. The resulting agglomeration schedule (the stages at which cases were combined based on the squared Euclidean distance between the cases) and the final dendogram generated are shown below.

According to Hair et al. (1995), the only assumptions that need to be checked in cluster analysis are representativeness and multicollinearity. Representativeness will be addressed in the limitations below. Multiple regression analyses were run to check the assumption that there is no multicollinearity. The VIF values generated were less than ten, indicating that this assumption was satisfied.

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	31	41	.021	0	0	16
2	4	38	.045	0	0	13
3	6	27	.049	0	0	7
4	13	44	.065	0	0	5
5	7	13	.068	0	4	23
6	10	25	.110	0	0	14
7	6	24	.129	3	0	18
8	39	40	.130	0	0	12
9	8	34	.149	0	0	27
10	17	30	.155	0	0	17
11	16	33	.160	0	0	25
12	15	39	.193	0	8	21
13	2	4	.206	0	2	28
14	10	28	.213	6	0	17

15	11	45	.231	0	0	22
16	29	31	.249	0	1	21
17	10	17	.270	14	10	19
18	1	6	.314	0	7	23
19	10	22	.332	17	0	27
20	36	37	.377	0	0	46
21	15	29	.407	12	16	29
22	9	11	.425	0	15	35
23	1	7	.425	18	5	30
24	3	21	.425	0	0	29
25	14	16	.438	0	11	36
26	20	23	.533	0	0	28
27	8	10	.596	9	19	30
28	2	20	.697	13	26	34
29	3	15	.747	24	21	37
30	1	8	.778	23	27	32
31	26	35	.869	0	0	33
32	1	32	1.154	30	0	37
33	18	26	1.305	0	31	40
34	2	5	1.327	28	0	41
35	9	19	1.331	22	0	38
36	14	42	1.488	25	0	40
37	1	3	1.749	32	29	42
38	9	43	2.274	35	0	44
39	46	47	2.527	0	0	43
40	14	18	2.955	36	33	41
41	2	14	3.610	34	40	42
42	1	2	3.343	37	41	44
43	12	46	3.902	0	39	45
44	1	9	4.220	42	38	45
45	1	12	6.108	44	43	46
46	1	36	13.749	45	20	0


Dendrogram using Centroid Linkage
Rescaled Distance Cluster Combine
0
5
10
15
20
25
Y
31 31
41 41
29 29
39 39
40 40
15 15
3 3
21 21
13 13
44 44
7 7
6 6
27 27
24 24
1 1
8 8
34 34
17 17
30 30
10 10
25 25
28 28
22 22
32 32
4 4
38 38
2 2
20 20
23 23
5 5
26 26
35 35
18 18
16 16
33 33
14 14
42 42
11 11
45 45
9 9
19 19
43 43
46 46
47 47
12 12
36 36
37 37

Discussion

According to Moisl (2015), items in a dendogram that have a shorter distance to their connecting node are more closely related than items with a longer distance to their node. For example, cases 36 and 37 have a much shorter distance to their node than do cases 46 and 47; this indicates that the vowels measured in 36 and 37 form a closer cluster than the vowels in cases 46 and 47.

The dendogram above shows that cases 36 and 37 form their own cluster, while the rest of the cases cluster together before clustering with the 36-37 cluster. Looking at the length values given in the Appendix, 36 and 37 are the longest vowels in the data set. This separation provides some support for the presence of both long and short vowels in Pahka'anil. Further research is needed, however, to provide stronger evidence.

Conclusion

As this was a pilot study, there were several limitations that need to be addressed when exploring this topic further. First, the measurements collected were only from one speaker. To increase the representativeness of the data, vowels from additional speakers should also be measured. Unfortunately, due to the status of the language this limitation can only be improved upon so much, but there are recordings from two other speakers that can be incorporated in the future.

Furthermore, there are many factors that need to be better controlled than they were in the current study. Stress patterns, location in the word, and the environment surrounding the vowel need to be controlled in the future.

Finally, the duration measurements were carried out by one person who knew the intent of the study. Although there was a protocol implemented for measuring the vowels, there still

may have been some experimenter bias. Ideally, future study would either have an additional person take the measurements and compare results, or would have another person rate the measurements taken by the experimenter.

This cluster analysis of Pahka'anil vowel length has tentatively supported the existence of both long and short vowels in Pahka'anil. As noted above, however, there are several limitations to this study which need to be addressed to provide stronger support for this claim.

References

Ahland, Michael. (2009). Aspects of Northern Mao (Bambassi-Diddesa) Phonology. *Linguistic Discovery*, *7*(1), 1-42.

Boersma, Paul & Weenink, David (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.24, retrieved 24 January 2017 from http://www.praat.org/

California Language Archive. (n.d.). Retrieved from http://cla.berkeley.edu/

Elicitation of numbers, LA 80.008, Berkeley Language Center, University of California, Berkeley, http://cla.berkeley.edu/item/15739[2]

Elicitation of words related to animals, LA 80.012, Berkeley Language Center, University of California, Berkeley, http://cla.berkeley.edu/item/15743

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis.* (4th ed. / Brian S. Everitt, Sabine Landau, Morven Leese. ed.). London: New York: Arnold; Oxford University Press.

Fulop, S., & Warren. (2014). An acoustic analysis of advanced tongue root harmony in Karajá. *The Journal of the Acoustical Society of America*, *135*(4), 2292.

Hair, J. F. J., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). Multivariate data analysis. Saddle River.

Jones, Caroline, Meakins, Felicity, & Muawiyath, Shujau. (2012). Learning Vowel Categories from Maternal Speech in Gurindji Kriol. *Language Learning*, *62*(4), 1052-1078.

Lamb, S. M. (1958). Linguistic prehistory in the Great Basin. *International Journal of American Linguistics*, *24*(2), 95-100.

Marean, L. (2015). Unpublished draft of Pahka'anil-English dictionary.

[2] Citations of the recordings used are given in the preferred citation format indicated by the California Language Archives.

Mary Chico, Sydney M. Lamb, Hansjakob Seiler. The Sydney M. Lamb and Hansjakob Seiler collection of Tübatulabal sound recordings, LA 80, Berkeley Language Center, University of California, Berkeley, http://cla.berkeley.edu/collection/10101[3]

Moisl, H. (2015). *Cluster analysis for corpus linguistics*. Retrieved from https://ebookcentral.proquest.com

Pagano, A. S., Figueredo, G. P., & Lukin, A. (2015). Measuring Proximity Between Source and Target Texts: an Exploratory Study. *Recent Contributions to Quantitative Linguistics*, *70*, 103.

Voegelin, C. (1935a). *Tübatulabal grammar* (University of California publications. American archaeology and ethnology; v. 34, no. 2). Berkeley, Calif.: University of California Press.

Voegelin, C. (1935b). *Tübatulabal texts* (University of California publications. American archaeology and ethnology; v. 34, no. 3). Berkeley, Calif.: University of California Press.

Voegelin, C. (1958). Working Dictionary of Tubatulabal. *International Journal of American Linguistics*, *24*(1), 221.

Wright, R., & Nichols, D. (2015, June 18). Measuring Vowel Duration. Retrieved from https://zeos.ling.washington.edu/~labwiki/w/index.php/Measuring_Vowel_Duration

[3] Preferred citation form provided for The Sydney M. Lamb and Hansjakob Seiler collection of Tubatulabal sound recordings.

Appendix

ID	Word	F1 (Hz)	F2 (Hz)	Length (s)
1	k**a**waiyo'	481.918	1215.162	0.0469
2	p**a**ga'	632.666	1433.984	0.1925
3	pag**a**'	506.473	1581.871	0.1381
4	p**a**ga'	608.031	1470.357	0.2203
5	pag**a**'	524.811	1631.734	0.19
6	t**o**xoil	523.635	1304.796	0.0674
7	t**o**xoil	483.692	1207.949	0.0824
8	ts**u**mil	388.907	1570.456	0.0739
9	tsum**i**l	443.261	2140.949	0.136
10	ts**u**mil	411.345	1376.389	0.1018
11	tsum**i**l	422.392	2209.903	0.1026
12	p**o**niu	711.889	1031.197	0.0974
13	p**o**niu	476.396	1122.607	0.0921
14	**u**nal	424.396	811.157	0.1858
15	un**a**l	563.04	1517.491	0.1288
16	**u**nal	410.198	907.211	0.2124
17	un**a**l	407.599	1510.94	0.1253
18	**i**ct	375.366	1817.709	0.2161
19	**i**ct	393.367	2314.777	0.1768
20	**a**cawɨt	644.686	1358.991	0.2716
21	ac**a**wɨt	509.668	1368.505	0.1152
22	acaw**ɨ**t	447.041	1520.431	0.0769
23	**a**cawɨt	687.999	1490.324	0.2361
24	ac**a**wɨt	491.845	1444.508	0.0704
25	acaw**ɨ**t	414.515	1406.759	0.0806
26	**a**cawɨt	375.779	1526.826	0.2599
27	ac**a**wɨt	514.355	1377.924	0.0617
28	acaw**ɨ**t	413.131	1227.009	0.1031
29	y**i**hawal	610.668	1664.199	0.0843
30	yih**a**wal	441.889	1458.158	0.1096
31	yihaw**a**l	630.309	1487.703	0.0964
32	k**u**yul	335.281	1154.566	0.0863
33	kuy**u**l	386.974	1030.963	0.2219
34	k**u**yul	344.848	1612.616	0.0691
35	kuy**u**l	375.902	1265.227	0.2176
36	tc**i**tc	379.153	2562.884	0.3005
37	tc**i**tc	373.201	2533.903	0.2605
38	n**a**nau	598.809	1521.697	0.2107

39	n**a**nau	610.733	1478.933	0.1419
40	n**a**pai	616.993	1454.251	0.1188
41	n**a**pai	615.091	1480.086	0.0918
42	n**o**'mdzin	489.723	1004.351	0.1417
43	no'mdz**i**n	568.765	2087.119	0.1789
44	n**o**'mdzin	458.962	1197.583	0.0974
45	no'mdz**i**n	406.184	2048.493	0.0896
46	n**a**nghan	745.292	1624.361	0.1325
47	nangh**a**n	901.804	1554.122	0.1925