

## README

This folder contains the primary data used in the research article:

Alderete, John and Kayleigh MacMillan. 2014. Reduplication in Hawaiian: Variations on a theme of minimal word. *Natural Language and Linguistic Theory* (anticipated online publication date, summer 2014).

Because of the large amount of work in collecting and coding this data, if the data is used in a future research publication, the authors believe it is appropriate to cite both the above work and the primary source below.

Pukui, Mary Kawena and Samuel H. Elbert. 1986. *Hawaiian dictionary*. Honolulu: University of Hawaii Press.

## DATABASE CODING – BASIC INFO ON ENTRIES

All of the fields in the database (i.e., the columns in the xls spreadsheet) are explained below. The column letter in the .xls document is given, together with the nickname for the field for all fields, e.g., Column F = ‘phonetic’, which is the phonetic form of the reduplicated word. Also, ‘RED’ = reduplicant, or copied part in a reduplicated word.

### Column A: record

A single word in the dictionary may have multiple entries, each one being a different sense of the same word. The spreadsheet will list each sense in a separate row. Each row (i.e., each word sense entry) has a unique identifier in Column A. An example of this identifier would be 'A.33.3' which corresponds to the 3rd word sense of the 33rd word in the dictionary under the letter A.

### Column B: ortho

The orthographic form of a word, as given in the dictionary.

### Column C: ppn

The reconstructed Proto-Polynesian form of the word, if given in the dictionary.

### Column D: gloss

The gloss associated with the sense of the word in A-B, as given in the dictionary. This was used in analyzing the meaning and other things like the morphological composition of the word.

### Column R: notes

Any comment on the coding decisions that are not easy to express with the fixed set of variables, e.g., odd morphological consistency or ambiguities in what the base or the reduplicant are.

## DATABASE CODING – MORPHOLOGY/SEMANTICS

Note: phrases are treated differently than words. Some reduplicated phrases and words arise from a sequence of two words, shown with a space between them in the base, e.g., /lele walo/ --> lele wa-walo. We analyze the attributes of the word within the phrase, not the entire phrase, because the reduplication process applies to a subpart of the phrase. Also, there are many forms in the definition that have other reduplicated words that are morphologically complex, especially with the prefix /ho'o-/. These are marked in the definition column with a [mcomplex, NameOfCoder] tag, so we can come back to them later.

#### Column E: base

The morphological base for the reduplicated word, as indicated in the dictionary entry.

#### Column I: basepos

Values: n, nvi, nvt, nvs, v, vi, vt, vs, num, pas/imp, interj, loc, unkn, ???

The part of speech of the base (=E) of the reduplicated word (=B, F). The part of speech tag is based on the categories used in the Hawaiian grammar (Elbert & Pukui 1979), which has important aspectual classes tied to the noun/verb distinction.

Nota bene: When a base has more than one Basepos, they will be listed with a hyphen in between them in the order they are listed in the dictionary. Eg nvt-vt. If the base has multiple definitions with the same Basepos, the Basepos is only listed once. If a base has two definitions that are marked as n and one as vt, the Basepos entry would be: n-vt. When the base is indicated as the same as another form or the variant of another form, the Basepos of that form is recorded, and a note is made in the notes column. When the base is a RED ('Red of Red'), then the Basepos is marked as unkn (unknown). When the entry for the base indicates to 'see' another form, then the Basepos is marked as unkn (unknown). This is because 'see' does not mean that the forms necessarily share a Basepos. When the entry for the base says that it is probably similar to another form, the Basepos is marked as unkn (unknown). This is because we cannot be sure that it is the same. When the base is said in the dictionary to be the probable base of the RED, the Basepos is marked as '???'. '???' is also used when there is no lexical entry for the base or it is morphologically complex. When there are two (possible) bases, the Basepos is marked as "?base1/base2" or "?X"(if both bases have the same Basepos). Thus, if the Basepos of base 1 is n-vt and the Basepos of base 2 is nvs, then the Basepos would be: ?n-vt/nvs.

#### Column N: semcat

Values: rep/cont, int, plur, dimin, speci, sim, same, unkn, other, conv, sub, gen, unfoc

The semantic categories associated with the meaning of the reduplicated words. The specific values shown above are explained below. If a RED has more than one Semcat both will be listed in the cell. This is determined by the presence of a semicolon (;) in the lexical entry. They will be separated by a hyphen (-) and placed in the order they appear in the dictionary entry. If a RED needs more than one Semcat to characterize one meaning in the dictionary entry, they will be separated by an ampersand (&); for

example, if the RED refers to an intensification of a subset of the base, then the semcat entry will be: sub&int.

Repetition/continuation (rep/cont): RED refers to a repeated or continuous action of the base.

Intensity (int): RED refers to the intensification of the attributes of the base.

Plurality (plur): RED refers to a plural number of the base.

Diminution (dimin): RED refers to a reduction or a smallness of the base.

Specificity (speci): RED refers to a specific or special type of the base.

Similarity (sim): RED refers to a thing or living being that has some of the properties that the base denotes.

Semantically same (same): RED refers to the same meaning as the base.

Unknown (unkn): RED refers to a case where the meaning of the RED is undeterminable either by itself or in relation to the base. For the most part, however, these are the unglossed reduplicated words, and the dictionary explains that these have the prototypical reduplicative meanings ‘frequentative, increased action, plural action’.

Other (other): RED refers to an idiosyncratic derivation that differs from that of the base.

Conversion (conv): RED refers to the changing of the Basepos of the base.

Subset (sub): RED refers to a subset of the meanings of the base.

General (gen): RED has a more general meaning than that of the base.

Unfocused (unfoc): RED seems to refer to an unfocused version of the action to which the base refers.

## DATABASE CODING – PHONOLOGICAL STRUCTURES

Column F: phonetic

A phonetic transcription of the orthographic form in B that marks syllable boundaries and stress, and uses phonetic symbols, like ʔ for the apostrophe. The stress and syllabification was generated automatically with a python algorithm, and we have noted a couple of cases where the algorithm placed stress in the wrong place. We have corrected these, but there may also be a couple additional cases.

Column G: basemora

An integer that records the number of moras in the base of the reduplicated word. This is useful for testing hypotheses that distinguish mora parity, like with foot-based analyses of the reduplicant.

Column H: v1len, short for 'vowel #1 length'

Values: short, long

'short' = the first vowel of the base is short, 'long' = the first vowel of the base is long.

Useful in studying length alternations.

Column J: edge, i.e., the position of the reduplicant with respect to the word edges

Values: l, r, either, both, mid

'l' = RED is leftmost in word (prefix), 'r' = RED is rightmost in word (suffix), 'either' = RED is either right or left when the RED is the full length of the base, both at the beginning and end (possible with multiple reduplication), 'mid' = RED is not a prefix or suffix, so it is an infix.

Nota bene: there are some cases where this is ambiguous for mid/r with syllable-sized REDs. Of the 25 potentially coded syll/r, 20 of them involve identical penultimate and final syllables in the base, so the redup form has three identical syllables in a row. Since there are so many more syllable infixes, and 80% of these cases have a specific shape, it makes sense to classify them as the more viable class of infixes, so they are 'mid'.

Column K: size, i.e., the size of the reduplicant in prosodic units

Value: syll, foot, more

'syll' = RED is a syllable, 'foot' = RED is exactly a foot (= two moras, given the analysis of stress and diphthongs), 'more' = RED is greater than two moras.

Column L: whole

Values: yes, no

'yes' = RED is identical to the base, so we have complete reduplication, 'no' = the base of reduplication has material that is not copied into the reduplication, so partial reduplication.

Column M: source, answers question, 'Where is the RED copied from?'

Values: begin, end, mid, beg/mid, either

'begin' = RED is copied from the beginning of the word, i.e., the first segment of the RED is the first segment of the base, etc., 'end' = RED is copied from the end of the word, 'mid' = RED is copied from somewhere in the middle of the word, like with syllable infixation, 'beg/mid' = it's unclear if 'beg' or 'mid', 'either' = copying could be from either the beginning or the end, as in complete reduplication of whole words,

Nota bene: if there is a choice between more than one source string, it is assumed that the string that is adjacent to the RED is the correct one. Example: kaamuumu, from kamumu; we assume that the penultimate syllable of the base form is the source, so source=mid. This assumption was guided by our knowledge that syllable infixing reduplication is pretty common.

Nota bene: in cases where two possible analyses exist, we apply Occum's Razor, and chose the simplest one. That is, if we have to chose between a foot reduplication with shortening of the base and a syllable reduplication (in the middle of the word), we chose the syllable reduplication. For example: kiipuu→kiipupuu, /pu/ is the RED as it does not involve any shortening of vowels. However, in cases where it is not possible to determine the base portion due to vowel and lengthening, neither portion is selected. That is, we indicate that the source and edge could be either and that the lengthening or shortening occurs in either the base or RED. Eg., maakuu→,maakumakuu, Edge: either, Source: either

Column O: basevlen, or 'vowel length alternations inside the base of reduplicated word'  
Values: short, long, ?short, ?long

'short' = the vowel length in the base portion of the reduplicated word is shortened relative to the base form, 'long' = if the vowel length is lengthened in the base portion of the reduplicated relative to the base form, '?long' or '?short' is used to indicate that it is unclear whether this occurs in the base or not. This is used when it isn't possible to determine what part is the base, or when there are two (possible) bases.

Column P: redvlen, or 'vowel length alternations in the RED, relative to the base'  
Values: short, long, ?short, ?long

'short' = the vowel length in the RED portion is shortened relative to the base form, 'long' = the vowel length in the RED portion is lengthened relative to the base form, '?long' or '?short' is used to indicate that it is unclear if whether this occurs in the RED portion or not (this is used when it isn't possible to determine what part is the base, or when there are two (possible) bases)

Nota bene: If either the base or the RED portion has more than one vowel length change, they will be separated by a hyphen (-), and placed in the order in which the vowels occur.

Column Q: cool, or 'unusual or interesting pattern'

This column contains a number of observations that are interesting, but not common enough to be registered in the other fields. 'dbl' = double reduplication, or coping of more than one identical syllable, e.g., hi-hi-hiki from hiki, 'mult' = multiple reduplications within the same word from different source syllables, e.g., ,pa.pa.,no.a.'no.a from panao. 'RED of RED' = the reduplicated word has a base that also appears to be the output of reduplication. Other things like deletion and English loans are noted.