LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 8 (2014), pp.738–777
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/ 24623

# The experimental state of mind in elicitation: illustrations from tonal fieldwork

Kristine M. Yu
*University of Massachusetts Amherst*

This paper illustrates how an "experimental state of mind", i.e. principles of experimental design, can inform hypothesis generation and testing in structured fieldwork elicitation. The application of these principles is demonstrated with case studies in toneme discovery. Pike's classic toneme discovery procedure is shown to be a special case of the application of experimental design. It is recast in two stages: (1) the inference of the hidden structure of tonemes based on unexplained variability in the pitch contour remaining, even after other sources of influence on the pitch contour are accounted for, and (2) the confirmation of systematic effects of hypothesized tonal classes on the pitch contour in elicitations structured to control for confounding variables that could obscure the relation between tonal classes and the pitch contour. Strategies for controlling the confounding variables, such as blocking and randomization, are discussed. The two stages are exemplified using data elicited from the early stages of toneme discovery in Kirikiri, a language of New Guinea.

**1. INTRODUCTION**[1] This paper revisits a very old method for studying tone languages: "how do you say X?" Summed up like this, elicitation from linguistic consultants might not seem to have the capacity to be particularly well-defined or rigorous as a method. But that would be a mischaracterization, as is evident from Pike (1948)'s brief summary of elicitation methodology for discovering tonemes in his classic tome on how to study a tone language:

> The procedure indicated is basically a method of controlling free, conditioned, key, mechanical, morphological, and sandhi tonal changes by inserting lists of words into selected contexts so as to reduce the number of variables at any one time and give the investigator the opportunity of observing the significant linguistic pitch in its simplest contrastive forms. (Pike 1948, p. vii)

Pike's description evokes an *experimental state of mind* in elicitation. By referring to the experimental state of mind, I take inspiration from Hyman (2001)'s conception of fieldwork as a state of mind: just as Hyman remarks that "it is possible to be a fieldworker without constantly going to the field," so is it possible to be an experimentalist in pursuing fieldwork without ever stepping foot in a lab. Bringing the lab to the field, e.g. conduct-

[1] I thank our Kirikiri consultant Alfius Polita for his generosity and patience as well as the other members of the Kirikiri team from the 2011 ANU tone workshop: Anthony Woodbury, Rebecca Hetherington, Peter Appleby, Rosey Billington, Lea Brown, Isolde Kappus, and Sutriani Narfafan, and we thank Steven Bird, Mark Donohue, Larry Hyman, and Mark Liberman for organizing the inspiring tone workshops. I also thank Brian Dillon and John Kingston for invaluable discussions and two anonymous reviewers for their suggestions. This paper is dedicated to Will Leben, my first mentor in linguistics.

E-ISSN 1934-5275

ing psycholinguistic experiments in the field, is a burgeoning line of research—see for instance, the forthcoming special issue of *Language, Cognition and Neuroscience* (Jaeger et al. to appear). However, the focus of this paper is to illustrate how an experimental state of mind can inform hypothesis generation and testing in traditional elicitation methodology in fieldwork. The theme of applying principles of experimental design in discovering tonal contrasts is pervasive throughout this volume.

This paper draws attention to the strong connections between labwork and fieldwork: it takes inspiration from Hyman (2007)'s remark that "elicitation is experimental phonology" and illustrates how elicitations can be structured in terms of principles of experimental design and analysis. The principles illustrated are applicable regardless of the linguistic phenomenon and data being elicited, but this paper focuses on: (1) the problem of discovering the tonemes of a language and (2) how to design structured elicitations to tackle this problem. I present illustrative examples from experiences in discovering the tones of Kirikiri (New Guinea, Lakes Plain) in fieldwork conducted with team members at the Prosodic Systems in New Guinea Workshop in December 2011. (Kirikiri words are given as preliminary broad IPA transcriptions; time did not allow for us to carefully work out segmental and length contrasts or stress or to do justice to the rich allophony due to pervasive lenition processes.)

The structure of the rest of the paper is as follows: it begins by using Pike's toneme discovery procedure as a starting point for illustrating the role of experimental design in elicitation. First, it introduces Pike's toneme discovery procedure in §2.1, and then recasts it in two stages in terms of principles of experimental design in §2.2 and §2.3. Section 2.2 describes the first stage of toneme discovery as uncovering unexplained variability in the pitch contour. From this evidence, we necessarily infer that the hidden structure of lexical tonal classes is present. Section 2.3 explicates the second stage of toneme discovery, at which point we have hypothesized tonal classes. In this stage, we work to confirm systematic effects of the hypothesized tonal classes on the pitch contour using elicitations carefully structured to control for confounding variables that could obscure those systematic effects. Sections 2.2 and 2.3 also include examples illustrating toneme discovery in terms of the principles of experimental design discussed from fieldwork in Kirikiri (§2.2.1, §2.3.3). Since the Kirikiri examples include work with fundamental frequency contours extracted from acoustic recordings rather than with perceived pitch contours, the paper includes a note on the relation between fundamental frequency and pitch in Box 1 at the beginning of the first Kirikiri example (§2.2.1). In addition, the materials used to prepare the Kirikiri examples can be found at http://www.krisyu.org/blog/posts/2013/10/preparing-data-for-ldc-paper/. Finally, §2.4 briefly exemplifies how the principles of experimental design discussed in the paper can be used to generalize beyond Pike's elicitation methods in tonal fieldwork in; it recasts Hyman (2007)'s elicitation work on tonotactic restrictions in Thlantlang Lai in terms of experimental design. Section 3 concludes.

In addition to the body of the paper presented here, a major component of the paper is a set of tutorials on practical aspects of collecting and analyzing phonetic data from elicitations. These tutorials and supporting files are referred to in the body of the paper and presented as supplementary material online at http://www.krisyu.org/blog/posts/2013/06/ldc-kiy-overview/.

**2 THE EXPERIMENTAL STATE OF MIND IN TONEME DISCOVERY.** This section walks through Pike (1948)'s toneme discovery algorithm (§2.1), shows how it can be conceived as an application of principles of experimental design in two stages (§2.2, §2.3), and then illustrates how these principles can be used to generalize elicitation design beyond Pike's discovery procedure in §2.4.

**2.1 PIKE (1948)'S TONEME DISCOVERY PROCEDURE: A WALKTHROUGH.** Pike's elicitation methodology for discovering tonemes in Pike (1948, Ch. 4, p. 48-54) consists of two main steps:[2]

1. Classification of words into word classes (*substitution lists*) of uniform morphological and phonological structure
2. Classification of words into tentative tonal classes by their pitch contours within controlled utterance contexts (*substitution frames)*

Figure 1 presents an overview of the entire two-step procedure, while Figure 2 illustrates the first step of classification of words into word classes in detail, and Figure 3 exemplifies the second step of examining the tonal properties of words in controlled contexts. Figure 1 illustrates that this second step of classifying words by their pitch contours within a particular substitution frame is repeated for each substitution frame for each substitution list compiled from the first step.

As Pike states, the purpose of the first classification of words by morphological and phonological properties is to control properties of the words such that any variation in pitch contour of words is highly likely to be due *only* to tonemic contrast:

> The first classification brings together words which are somewhat alike in phonetic and grammatical structure. Such a grouping tends to reduce the hazards introduced in the analysis of these words by segments which cause nonphonemic modification of tonemes. The ear is distracted in its listening for pitch when the forms of the items under attention are not comparable. (Pike, 1948, p. 48)

The morphological and phonological properties listed for consideration in this first classification step in Figure 2 (lexical class, length, syllable shape, etc.) should not be taken to be a rigid prescription for exactly how to partition words into classes for pitch contour comparison. The set of properties listed may be neither necessary nor sufficient for toneme discovery in a particular language. Moreover, the output word classes, the *substitution lists* from the classification, may well be adjusted in the course of fieldwork to yield a coarser or finer-grained partition of the words. For instance, the classification tree in Figure 2 currently does not distinguish vowel quality, but in the course of tonal fieldwork, if one suspected an interaction between vowel height and pitch contours, a natural step would be to include vowel height in word classification. This heightened attention to vowel quality in classification would yield a finer-grained partition of the words and introduce an additional level of depth to the classification tree.

[2] Pike further describes methods for determining the number of tonemes and their phonological description in Ch. 5, but we focus on his first two steps for the purposes of illustrating the experimental state of mind in elicitation in this paper.

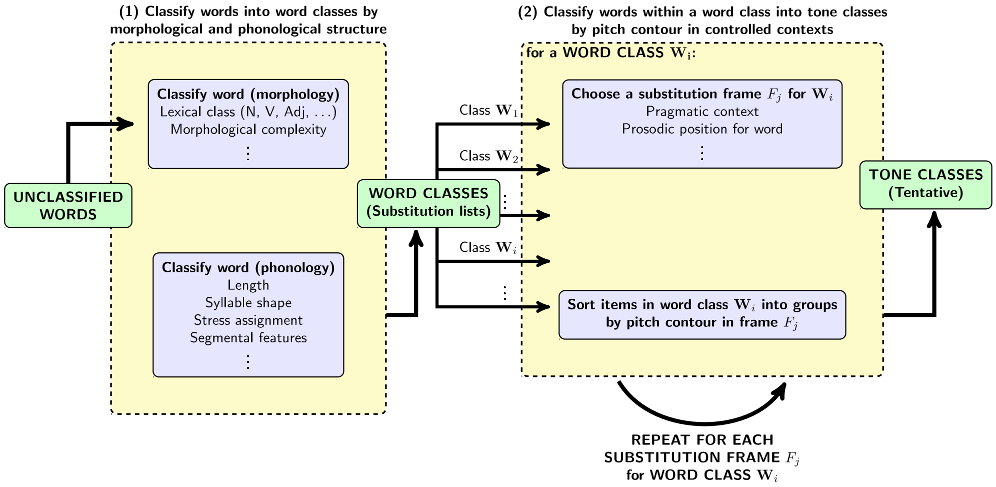

(1) Classify words into word classes by morphological and phonological structure
(2) Classify words within a word class into tone classes by pitch contour in controlled contexts
for a WORD CLASS $\mathrm{W}_i$:
UNCLASSIFIED WORDS
Classify word (morphology)
Lexical class (N, V, Adj, ...)
Morphological complexity
Classify word (phonology)
Length
Syllable shape
Stress assignment
Segmental features
WORD CLASSES (Substitution lists)
Class $\mathrm{W}_1$
Class $\mathrm{W}_2$
Class $\mathrm{W}_i$
Choose a substitution frame $F_j$ for $\mathrm{W}_i$
Pragmatic context
Prosodic position for word
Sort items in word class $\mathrm{W}_i$ into groups by pitch contour in frame $F_j$
TONE CLASSES (Tentative)
REPEAT FOR EACH SUBSTITUTION FRAME $F_j$ for WORD CLASS $\mathrm{W}_i$

FIGURE 1: A schematic diagram summarizing the discovery procedure for tonemes in Pike (1948, Ch. 4). There are two main steps: (1) classification of words by morphological and phonological structure into word classes (substitution lists), and for each substitution list, (2) classification of the items in the list into tentative tonal classes by their pitch contours in a controlled utterance context, a substitution frame. For all substitution lists, the second step is repeated for all applicable substitution frames for the substitution list. Proposed tonal classes may be adjusted after each iteration. Abbreviations: N = noun, V = verb, Adj = adjective.

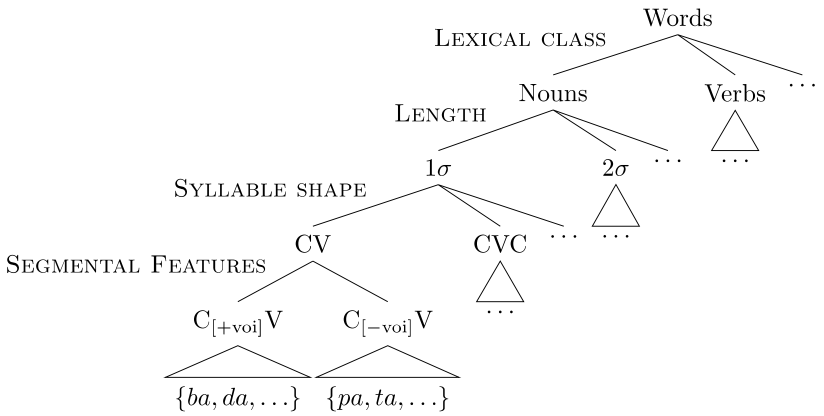


FIGURE 2: A classification tree depicting the initial step of partitioning words by morphological and phonological structure in the discovery procedure for tonemes in Pike (1948, Ch. 4). Words are partitioned by lexical class, length (in syllables), syllable shape, and segmental features (here, onset voicing). The sets of terminal elements {*ba*, *da*, ...}, {*pa*, *ta*, ...} are word classes Pike calls *substitution lists*. These word classes are uniform in morphological and phonological structure and are to be elicited in controlled contexts in the second step of the discovery procedure.

The output of the first classification step, the assignment of words to word classes (*substitution lists*), is the input to the second classification step schematized in Figure 3. For each substitution list, a set of *substitution frames* is generated, as shown in the middle box labeled "substitution frames" in Figure 3. The specification of these frames includes phrasal-level prosodic structure as well as pragmatic context.[3] For each iteration of eliciting a substitution list in the context of a substitution frame, the items in the substitution list are sorted into tentative tonal classes by their particular pitch contours in the substitution frame. For instance, in Figure 3, two words in substitution list $\boldsymbol{W}_i$, *naβi* 'dog' and *kaza* 'gecko', both have rising pitch contours in frame $F_j$, distinct from the pitch contours of other words in $\boldsymbol{W}_i$. Thus, these words would be sorted into a tentative tonal class together. Comparison of hypotheses for tonal groupings from different substitution frames for a substitution list provides converging or diverging evidence for particular proposed tonemes. These hypotheses may eventually also be compared to generalize not only over substitution frames within a substitution list, but also across substitution lists.

A final important point about the tonal classification step is the emphasis on the assignment of group labels to words rather than the nature of the labels themselves. Pike writes that: "up to this point there has been no essential need for tonal transcription. It is the grouping as such which has been important" (Pike, 1948, p. 55): in Figure 3, what's critical in the rightmost "pitch contours" box is not the IPA tonal transcription, but the colors of the boxed tones, which indicate group membership.[4]

The output of Pike's second toneme discovery step can be thought of as a list of mappings from words to glass jars in a pantry storing jars of words. There might be faded written labels on the jar lids, but what effectively serves as the jar labels in the pantry is the set of identifying and distinguishing properties of the contents inside a jar. Indulging a bit with this pantry metaphor: taking a shelf storing various kinds of pasta—we might not know whether to label a jar as *gemelli* or *bucatini*, but it's easy enough to keep the different pasta jars straight and to glance at the jars and distinguish one with little spirals in it from one with long tubes in it.

[3] Pike (1948, p. 51) actually discusses controlling "emotional context… to prevent intonational changes". We take "emotional context" to have a meaning roughly equivalent to that of *pragmatic context,* as a cover term for situational context that may interact with prosody, and to also include paralinguistic influences, such as degree of emphasis and modulation of attention (see Table B.1 in the Appendix).

[4] This observation about the importance of sorting and clustering pitch contours in contrast to the negligible role of the choice of representation of pitch contours in the early stages of toneme discovery was also discussed at the Berkeley Tone Workshop in 2011 (NSF Project Prosodic Systems in New Guinea: Integrating computational and typological approaches to linguistic analysis). It was the key motivation for the development of Toney, software to aid this process (See Bird and Lee, this volume).

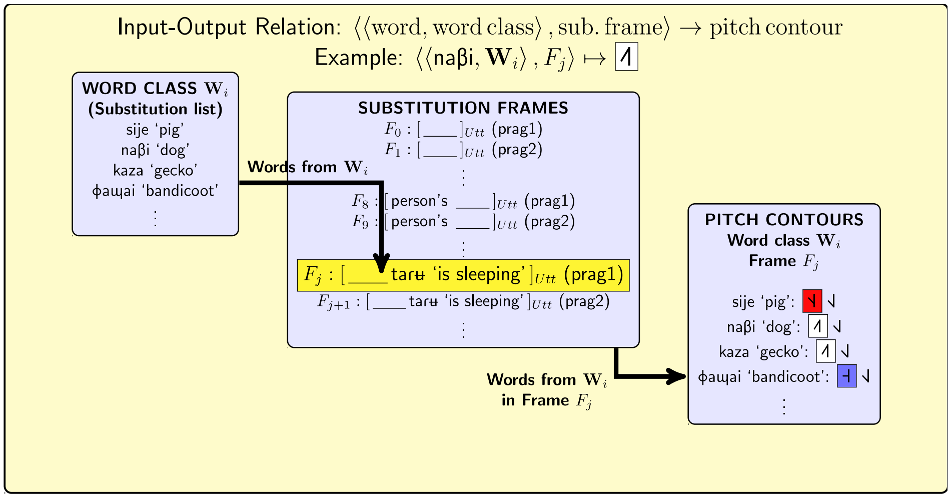

Input-Output Relation: $\langle\langle \text{word, word class}\rangle, \text{sub. frame}\rangle \rightarrow \text{pitch contour}$
Example: $\langle\langle$naβi, $\mathbf{W}_i\rangle, F_j\rangle \mapsto$ ˩˥˩
WORD CLASS $\mathbf{W}_i$
(Substitution list)
sije 'pig'
naβi 'dog'
kaza 'gecko'
ɸaɰai 'bandicoot'
⋮
Words from $\mathbf{W}_i$
SUBSTITUTION FRAMES
$F_0$ : [____]$_{Utt}$ (prag1)
$F_1$ : [____]$_{Utt}$ (prag2)
⋮
$F_8$ : [ person's ____]$_{Utt}$ (prag1)
$F_9$ : [ person's ____]$_{Utt}$ (prag2)
⋮
$F_j$ : [____ tarʉ 'is sleeping' ]$_{Utt}$ (prag1)
$F_{j+1}$ : [____ tarʉ 'is sleeping' ]$_{Utt}$ (prag2)
⋮
Words from $\mathbf{W}_i$ in Frame $F_j$
PITCH CONTOURS
Word class $\mathbf{W}_i$
Frame $F_j$
sije 'pig': ˦˨ ˦˩
naβi 'dog': ˩˥˩ ˦˩
kaza 'gecko': ˩˥˩ ˦˩
ɸaɰai 'bandicoot': ˧ ˦˩
⋮

FIGURE 3: A schematic diagram summarizing the second step in the discovery procedure for tonemes in Pike (1948, Ch. 4) using an example from Kirikiri (see §2.2.1). This consists of eliciting items in a word class (substitution list) $\boldsymbol{W}_i$ in substitution frames and sorting the items by pitch contour in a substitution frame into tentative tonal classes (tonemes). Among the substitution frames licit for $\boldsymbol{W}_i$, Frame $F_j$ positions items utterance-initially in a particular pragmatic context *prag1*. The elicited pitch contours (transcribed with IPA tone symbols) in frame $F_j$ for the words in $\boldsymbol{W}_i$ are clustered into tentative tonal classes. Here, *naβi* 'dog' and *kaza* 'gecko' are proposed as members of a tentative tonal class (colored white) since they have very similar pitch contours in context $F_j$ which are distinct from the pitch contours of other words in $F_j$, $\boldsymbol{W}_i$.

Having laid out Pike's toneme discovery procedure, I recast it in two stages from the perspective of experimental design: (1) treating tonal class as a *latent variable*, as a variable whose presence is to be inferred from the unexplained variability in pitch contours over words (§2.2), and (2) explicitly manipulating (putative) tonal classes as *independent variables* in testing hypotheses about the partition of words into tonal classes (§2.3). *Independent variables* refer to manipulations of the *experimental unit*—each elicitation item constructed and elicited by the fieldworker and uttered by the consultant—not manipulations of words in the language of study. When we discuss manipulating SYLLABLE SHAPE or TONAL CLASS as independent variables, we are not claiming to manipulate the syllable shape or tonal class of a word. Instead, we mean manipulating the properties of an elicitation item—properties of the substitution frame, e.g., the choice of pragmatic context and syntactic construction, as well as properties of the target word, e.g. the choice of the tonal class of the target word within the elicitation item.[5]

[5] Throughout this paper, reference to any independent or dependent variable is intended to be to the variable associated with the experimental unit of the elicitation item, e.g. "the pitch contour over a word" is short for "the pitch contour over a (target) word embedded in an elicitation item uttered by the consultant in an elicitation session."

Running illustrations of each stage from fieldwork on Kirikiri are provided in §2.2.1 and §2.3.3.

**2.2 EXPERIMENTAL DESIGN IN EARLY TONEME DISCOVERY: TONAL CLASS AS A LATENT VARIABLE.** At the earliest stages of studying a language, one may not be even sure whether it has lexical tonal classes, and even if one is confident that there are tonal classes in the language, one doesn't know enough about them to manipulate TONAL CLASS as an *independent variable* and explicitly group words into different tonal classes. The research hypothesis at this point is:

Hypothesis 1 (Existence of tonal classes) *There are tonal classes in the language of study.*

Note that tonal classes are not directly observable: one can only indirectly observe the presence of tonal classes through the *dependent variable* of the pitch contour over the word, which one observes as one elicits each new item with one's consultant. As is implicit in Pike's procedure, in order to proceed in discovering hidden structure which is not directly observable, one must make assumptions linking properties of speech that one can observe—*observed variables*—to unobservable, underlying tonal classes—*latent variables*. Such assumptions can be encoded in a *linking hypothesis*, which makes explicit assumptions about the relation between observed variables, e.g. pitch contours over words, and (unobserved) latent variables, e.g. tonal classes:[6]

Hypothesis 2: (Linking hypothesis between lexical tonal classes and pitch contours)

*Lexical tonal classes (tonemes) induce systematic variation in the pitch contours of words. Therefore, the unobservable concept of a toneme is observable via its influence on the pitch contours of words: if two words have sufficiently*[7] *different pitch contours, then they belong to different lexical tonal classes*.[8]

The linking hypothesis in Hypothesis 2 should make the reader sputter with incredulity: what about the many other potential sources of influence on the pitch contours of words

[6] Some of the earliest discussions of linking hypotheses in studying human behavior comes from the study of human vision (Brindley 1960, Teller 1984), in which linking hypotheses must be made about the mapping between perceptual and psychological states. One can also view the situation here as an instance of making linking hypotheses about the mapping between perceptual (auditory) states (pitch contours over words) and psychological (cognitive) states (tonal concepts—tonemes).

[7] I return to the notion of *sufficiently different* pitch contours later in the restatement of Hypothesis 2 as Hypothesis 2' (§2.3). For now, the reader can get an intuition of what is meant by glancing at Figure 8 (§2.3). Within each small box labeled with Tone 2, we see that the pitch contours show some variability. However, the variability within a tonal class is much smaller than the variability across tonal classes, between each Tone 2 box and the Tone 3 box directly below it.

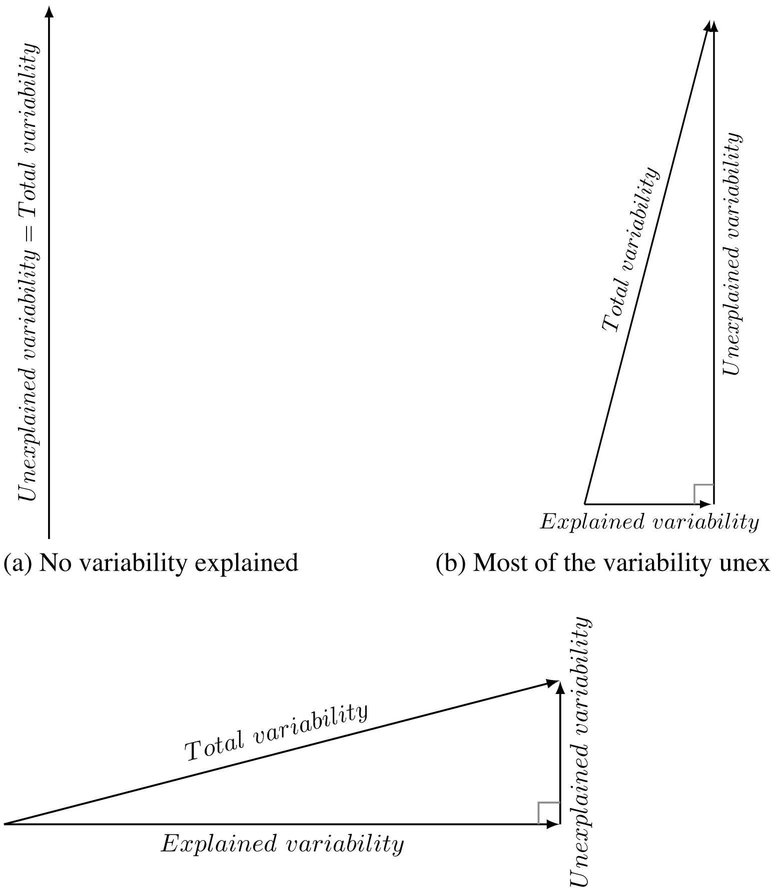
[8] There are acoustic and perceptual dimensions of tone other than those directly related to fundamental frequency (see Box 1) and pitch, such as amplitude, e.g. see Whalen and Xu (1992), duration, e.g. see Howie (1976), and phonation, e.g. see Belotel-Grenié and Grenié (1997) and Andruski (2006), but a discussion of these is beyond the scope of this paper.

uttered during an elicitation that were discussed in §2.1? (In fact, there are many, many more additional sources of influence on pitch not discussed in §2.1, including physiological, grammatical, and sociolinguistic ones, a sampling of which are listed in Table B.1 in the Appendix. (Some of these are not in play within a speaker within an elicitation session.) See also Snider, this volume, for further discussion of potential confounds in establishing tonal contrast.) Suppose, for instance, that we elicited one word at the end of the utterance and another one in the middle of the utterance. If these two words were uttered with different pitch contours, and we concluded on the basis of Hypothesis 2 that the two words belonged to different tonal classes, we could be mistaken: perhaps the pitch differences were actually due to the difference in prosodic position between the two elicited utterances.

There is still a sense in which Hypothesis 2 is reasonable, though. (LEXICAL) TONAL CLASS could certainly be *one* variable in a model explaining variability in the pitch contour over a word, and PROSODIC POSITION and any other of the variables mentioned in §2.1 could be *explanatory variables* as well—independent variables that are central to our research questions and hypotheses. There could be *many* explanatory variables in a model of pitch contour variation. In order to provide evidence that TONAL CLASS is one of these explanatory variables (Hypotheses 1 and 2), we follow the strategy below:

- Propose a set of independent variables to include in a model to explain pitch contour variation. This set does not include TONAL CLASS, since we are not yet at the stage where we can manipulate TONAL CLASS as an independent variable.
- See how far the set of explanatory variables in the current model goes towards explaining the variability in pitch contours of words from elicited utterances.
  - If the leftover unexplained variability is huge,[9] we suspect that we may have missed one or more important explanatory variables in our model, and if we have considered a sufficiently wide range of possible variables for our model, then we can conclude with some confidence that we need to have TONAL CLASS in the model to help explain the pitch contour variability (see Figure 4b). We can then proceed to check if there is a large gain in explained variability with TONAL CLASS in the model.
  - If the leftover unexplained variability is small, we have no compelling positive evidence for the language having lexical tonal classes. The small proportion of variability left unexplained is likely due to a constellation of secondary influences on pitch contour variation we've abstracted away from, e.g. perhaps some of those listed in Table B.1 in the Appendix (see Figure 4c).

[9] In statistical data analysis, what counts as "huge" is precisely defined in terms of the ratio of unexplained variance to total variance, where variance is a statistical estimate of variability about the mean; see Johnson (2008, §4.1), Rietveld and van Hout (2005, Ch. 4) and Maxwell and Delaney (2004, Ch.3) for introductions to analysis of variance. For application in fieldwork elicitation, an intuitive definition of "huge" is sufficient.


$Unexplained\ variability = Total\ variability$
$Total\ variability$
$Unexplained\ variability$
$Explained\ variability$
$Total\ variability$
$Explained\ variability$
$Unexplained\ variability$

(a) No variability explained

(b) Most of the variability unexplained

(c) Most of the variability explained

Figure 4: A geometric summary of partitioning the total variability in the pitch contour over a word into two parts (Saville and Wood, 1986): (1) the variability explained, i.e., the variability accounted for by the independent variables, e.g. those given in Table 1 and (2) the residual (unexplained) variability, which must be attributed to the influence of other variables, e.g. those not enumerated in Table 1—in particular, we hypothesize, the influence of the latent variable tonal class.

Under this strategy, we begin with a morass of variability in the pitch contour over a word (Figure 4a). None of this variability is explained, i.e. all of the variability is *unexplained*. Upon introducing variables into a model of pitch contour variation, we hope to be able to carve off some of that variability into *explained variability*, variability which is accounted for by our introduced explanatory variables. The better we understand what influences pitch contour variability, the more variability from the total variability we are able to partition into *explained variability*, leaving less *unexplained variability* remaining. If our model does not do a good job of covering the sources of influence for pitch contour variability, the partition between explained and unexplained variability looks like the breakdown in Figure 4b, where most of the variability remains unexplained. If our model does in fact do a good job, then the partition looks like the decomposition in Figure 4c, where most of the variability is explained by the proposed explanatory variables.

At this point, our guiding research question is:

Question 1: (Explaining variability in pitch contours) *Can we explain part of the variability in the pitch contour over a word in an elicited utterance? How much of this variability can we explain?*

Which explanatory variables should we include in a model of pitch contour variability? The answer from Pike's procedure is Hypothesis 3—the variables to include are in the list of variables we discussed in §2.1, which we summarize in Table 1. Unlike TONAL CLASS, we know enough about the variables in Table 1 to manipulate them as independent variables. For instance, we can explicitly manipulate the LENGTH of the target word in an elicitation item to be set at different *levels*—possible instantiations of a variable in the experimental design, e.g. monosyllabic, disyllabic, etc.

Hypothesis 3: (Explaining variation in pitch contours over words) *The primary influences on variation in pitch contours over words are the variables listed in Table 1.*

| Variable | Examples of levels |
|---|---|
| Lexical class | Noun, verb |
| Morphological complexity | Simplex, complex |
| Length | Monosyllabic, disyllabic |
| Syllable shape | CV, CVC, CCV, CV: |
| (Vowel length)[10] | (short, long) |
| Segmental features | Initial voiced stops, initial voiceless stops |
| Stress | Present, absent |
| Stress assignment | Initial, peninitial, 3rd syllable, penultimate |
| Prosodic position | Utterance-initial, isolation |
| Pragmatic context | Out-of-the-blue focus, contrastive focus on first NP |
| Syntactic structure | Possessive prenominal phrase, relative clause |

TABLE 1: Independent variables in Pike (1948)'s toneme discovery procedure. The table gives each variable name, and some examples of levels—possible instantiations of a variable in the experimental design, e.g. LEXICAL CLASS = noun; LEXICAL CLASS = verb. A horizontal line divides variables pertaining to the target word (items in the substitution lists) and variables pertaining to the context (substitution frame). Vowel length is listed in parentheses since Pike mentions it explicitly, but it could also be subsumed under syllable shape.

[10] See Remijsen, this volume, for a detailed treatment on investigating lexical tone when there are interactions between tone and quantity contrasts of vowel length and syllable shape and stress.

The small list of variables in Table 1 certainly does not exhaustively capture all aspects of the elicitation context in toneme discovery that may influence pitch contour variation over a word, since context is always unbounded. (Is the consultant lethargic or excited? How far away are you sitting from the consultant? What were the last five words elicited—has the consultant started building a discourse context around them? What time of day is it? How much sleep did the consultant get last night?). However, we can try to capture the primary aspects of relevant context: we can tackle contextual variables that we know of and that we suspect may account for a large amount of variability in the pitch contour of a word in an elicited utterance; we abstract from others.[11] There will always remain some unexplained variability.[12]

In sum, the initial stage in Pike's toneme discovery procedure can be cast in terms of seeing how far we can get in explaining variability in pitch contours of words without appealing to the hidden structure of tonal classes. We acknowledge every known source of influence, or at least all known primary sources of influence, on variability in the observable dependent variable, in an attempt to explain away all of the variability in the dependent variable, the pitch contour over the target word. Any residual unexplained variability in the dependent variable must be due to a set of variables that has been overlooked. If there is a large amount of unexplained variability remaining, we hypothesize that we are failing to take into account the latent variable TONAL CLASS.

**2.2.1 EXAMPLE: THE EARLIEST STAGES IN DISCOVERING KIRIKIRI TONES.** This section illustrates the earliest stages of toneme discovery, when TONAL CLASS is treated as a latent variable in fieldwork on Kirikiri. The data comes from the three earliest recorded elicitation sessions, 20111207-1-kiy-ap-wordlist, 20111207-2-kiy-ap-framedwordlist, and especially 20111208-6-kiy-ap-nps-vps: it is used to demonstrate the process of exploring potential sources of unexplained variability in the dependent variable of the pitch contour over the word. The Kirikiri datasets and their associated sound files are named following the convention [date]-[session_number]-[language_code]-[speaker]-[elicitation-description]. Further discussion of organizing data and files from elicitation can be found in the tutorial at: http://www.krisyu.org/blog/posts/2013/06/organizing-elicitation-items/.

Up to this point, the paper has taken pitch contour to be the dependent variable. In the illustrations of Kirikiri throughout this paper, though, the dependent variable is taken to be fundamental frequency ($f_0$) contour rather than the pitch contour. Fundamental frequency is the primary acoustic correlate of the auditory percept of pitch and can be extracted from audio recordings. For this paper, recordings were analyzed with Praat, free and open source

[11] The strategy of making methodological abstractions in the process of groping toward scientific understanding—setting some contextual factors aside to home in on what is of primary interest—is an old one. Plato (360 B.C.E) described it as carving nature at its joints in *Phaedrus* 265e.

[12] It is typical in psycholinguistic studies for most of the variability in the dependent variable (which is frequently reaction time in a subject's response to some stimulus) to be due to variability among subjects and items: the explanatory power from the variables of interest is swamped by subject-to-subject and item-to-item variability. Baayen (2008, p. 281–282) gives a characteristic example where only 0.3% of the total variability is explained by the independent variables.

software for speech analysis (Boersma and Weenink 2010), and $f_0$ was extracted using the RAPT algorithm (Talkin 1995). See Box 1 for a note on the relation between fundamental frequency and pitch and the tutorial at http://www.krisyu.org/blog/posts/2013/10/working-with-$f_0$-contours/ for further information about working with $f_0$ contours.

At the initial stages of tonal fieldwork, we begin with all the variability in the $f_0$ contours unexplained, the initial state illustrated in Figure 4a. A visualization of this chaotic state is shown in Figure 5, a plot of $f_0$ contours over target words (substitution items) for each of the recorded elicitation items from elicitation sessions 20111207-1-kiy-ap-wordlist and 20111207-2-kiy-ap-framedwordlist. In these sessions, each target word was elicited in isolation.

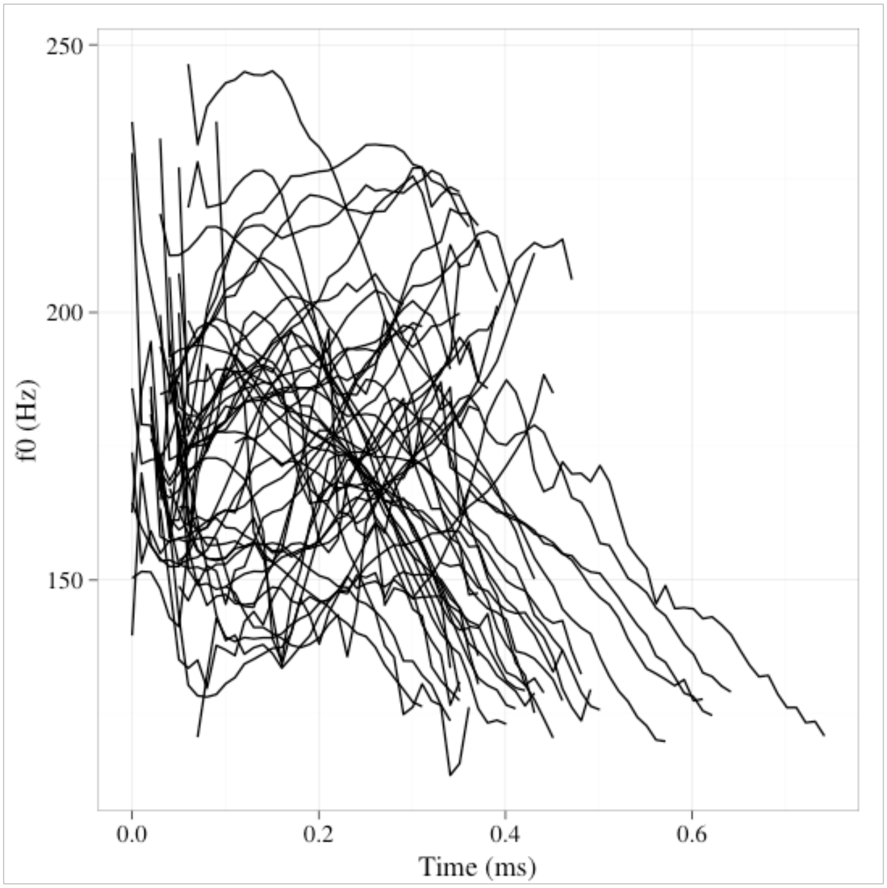


FIGURE 5: Fundamental frequency ($f_0$) contours of each of the recorded elicitation items from 20111207-1-kiy-ap-wordlist and 20111207-2-kiy-ap-framedwordlist.

**BOX 1 A NOTE ON THE RELATION BETWEEN FUNDAMENTAL FREQUENCY AND PITCH.** Fundamental frequency and pitch are often used interchangeably, but if we are being precise about terminology, they are distinct terms. *Pitch* refers to an auditory percept closely related to the rate of vocal fold vibration, or equivalently, the glottal pulse rate, which is manifested as the rate of repetition of cycles in the acoustic waveform—it is this rate of repetition that is called the *fundamental frequency*, $f_0$.[13]

Pitch is not something that can be directly extracted from a recording of speech because it is an interpretation of speech sounds that is mediated by the auditory pathways of the nervous system. What we *can* extract from recorded speech is fundamental frequency, as it is an acoustic parameter of the speech signal. Although we might refer to the timecourse of fundamental frequency estimated from a recording as the *pitch contour*, this is strictly a colloquial term for what is more accurately referred to as the *fundamental frequency contour* or $f_0$ *contour*. When we write down tonal transcriptions during an elicitation session, we are transcribing the perceived pitch contour, but when we examine the output of an $f_0$ estimation algorithm on a computer, we're examining the $f_0$ contour extracted from the speech signal, not the pitch contour.

The time course of fundamental frequency over a word is certainly informative about the time course of pitch over a word and vice versa, but the relation between the two is not transparent. First, language experience tunes the way the auditory system processes $f_0$ (Mattock and Burnham 2006, Mattock et al. 2008, Krishnan et al. 2005, Krishnan and Gandour 2009). The implication of language-dependent pitch perception for tonal fieldwork is that fieldworkers cannot assume that their perception of pitch in an elicited utterance is representative of the consultant's or anyone else's perception. Second, in addition to pitch being in the ear of the beholder, pitch perception is also integrated with perception of other qualities of the speech signal than $f_0$, such as intensity and other spectral properties (Baken and Orlikoff 2000, p. 146), as well as segmental context (Rose 1988, House 1990) and possibly other aspects of voice quality such as phonation, e.g. creakiness and breathiness (Sun and Xu 2002). Thus, although we can extract $f_0$ from the speech signal independently from other acoustic parameters, we cannot factor out pitch perception in speech independently from the concurrent perception of other properties of the speech signal, since perception of these properties is integrated.

Moreover, while we can estimate $f_0$ at a particular instant, pitch perception at any particular moment is strongly influenced by preceding context (Wong and Diehl 2003). That is, we can measure $f_0$ at a particular time, without reference to the past or future. But we cannot understand the perception of pitch in speech at some point in time without taking into account other information from the past (and possibly also the future) that provides a context for the perception of pitch at the current instant.

For more information on $f_0$ and pitch, see Baken and Orlikoff (2000, Ch. 6) and Moore

[13] Sometimes $f_0$ is capitalized as F0, but keeping it lower case helps reinforce the fact that the physics of fundamental frequency has very little to do with the physics of *formants*, vocal tract resonances critical in the description of vowel quality, which are standardly abbreviated as F1, F2, etc. for first formant, second formant...

(2012, Ch. 6). For more on working with $f_0$ contours, see Gussenhoven (2004, Ch. 1) and Ladefoged (2003, Ch. 4) and the tutorial at http://www.krisyu.org/blog/posts/2013/10/working-with-$f_0$-contours/.

With the data from 20111208-6-kiy-ap-nps-vps (see Table A.1 in the Appendix for the full list of elicitation items), we also begin with all the variability in the pitch contour over the word unexplained. In this elicitation session, the target words were common animal names (Table 2), and they were elicited in noun phrases (noun-adjective word order) and sentences with VPs (Table 3). The plot of $f_0$ contours for all substitution items for this elicitation session in Figure 6 is as chaotic as Figure 5.

To work towards explaining some of the variability, we begin to partition the $f_0$ contours by substitution frame and properties of the substitution item (target word) given in Table 2, such as length and lexical class. Across the substitution frames, prosodic position is fixed to be initial, and syntactic structure varies between (modified) NPs and VPs.

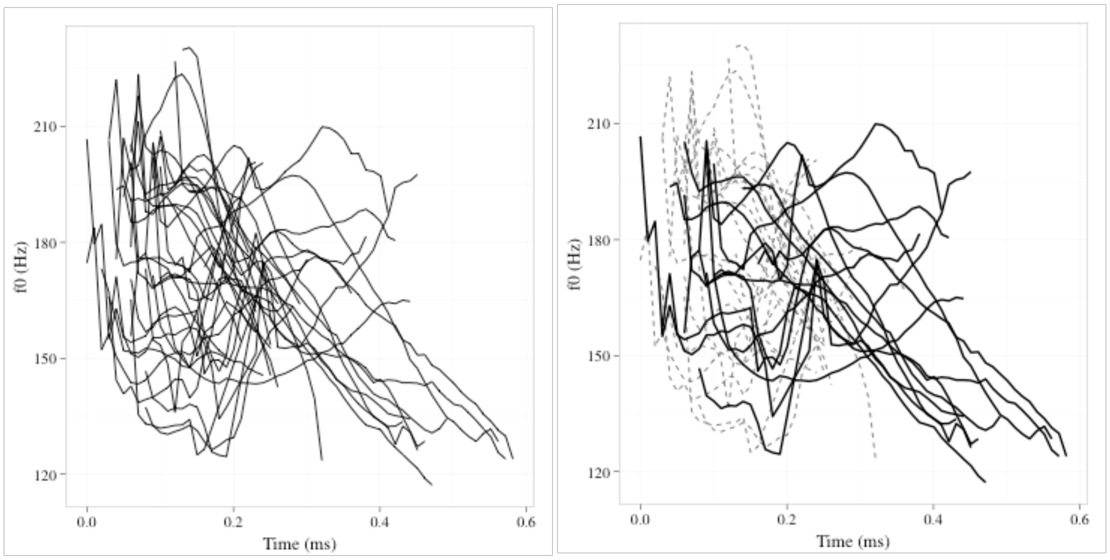
| Target word | Gloss | Lexical class | Length (syll) |
|---|---|---|---|
| kɔɔ | ant | noun | 1 |
| fɔɔ | wallaby | noun | 1 |
| sije | pig | noun | 2 |
| naβi | dog | noun | 2 |
| Kaza | gecko | noun | 2 |
| ɸaɰai | bandicoot | noun | 2 |

Table 2: A partial list of target words (substitution items) from 20111208-6-kiy-ap-nps-vps and some of their properties

| Substitution context | Kirikiri | Gloss |
|---|---|---|
| Isolation | # ___ # | |
| Adjective-black | ___ kɔɔ | black ___ |
| Adjective-small | ___ sɔɔ | small ___ |
| Adjective-female | ___ kuu | female ___ |
| VP-sleep | ___ taɾʉ | ___ is sleeping |
| VP-sound | ___ kʍaa zaɾe | ___ is making a sound |

Table 3: A list of substitution frames from 20111208-6-kiy-ap-nps-vps.

One immediate source of variability becomes clear when we differentiate between $f_0$ contours uttered in isolation and all other $f_0$ contours in the plots. Figure 6b shows that the $f_0$ contours uttered in isolation, drawn as solid black lines, have durations roughly twice as long as $f_0$ contours uttered in non-isolation contexts, drawn as gray dashed lines.



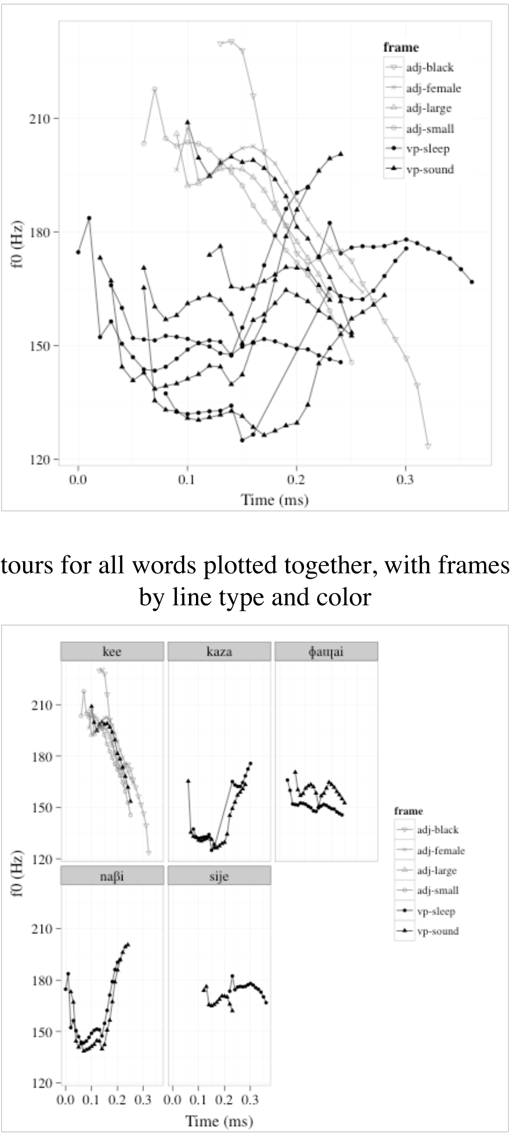
(a) All contours

(b) Contours from the isolation frame drawn in black solid lines; all other contours drawn in gray dashed lines

FIGURE 6: $f_0$ contours of target words from all recorded elicitation items in 20111208-6-kiy-ap-nps-vps. When the contours from the isolation frames are differentiated from contours from other frames in Figure 6b (on the right), we move some of the variability in the $f_0$ contours from being unexplained in Figure 6a (on the left) to being explained.

If we take a closer look at the non-isolation contexts, we find that further differentiation between substitution frames in plotting $f_0$ contours appears to reveal some additional structure in the set of $f_0$ contours. Figure 7a shows $f_0$ contours from non-isolation contexts, grouped by substitution frame (see Table 3). The $f_0$ contours from VP frames (black) include rises, while the $f_0$ contours from adjective frames (grey) are falls.

However, the relation between syntactic properties of the frame and the $f_0$ contour is likely spurious. In Figure 7b, we plot $f_0$ contours for each target word in separate subplots, and the $f_0$ contours within each subplot are very similar to one another despite being elicited in different substitution frames. It's clear from Figure 7b that the structural regularity in the $f_0$ contours we found in Figure 7a is not due to the effect of different substitution frames, but rather, due to the effect of the length of the target word: the $f_0$ contours for the monosyllabic *fɔɔ* and *kee* only include steep falls, while the $f_0$ contours for the disyllabic words do not. It was an accident that only monosyllabic target words were elicited in adjective substitution frames, while disyllabic as well as monosyllabic words were elicited in VP frames.



(a) $f_0$ contours for all words plotted together, with frames distinguished by line type and color

(b) $f_0$ contours plotted separately for each target word, with frame indicated by color and line type.

FIGURE 7: $f_0$ contours of the substitution items 'wallaby', *kee* 'ant', *kaza* 'gecko', *naβi* 'dog', *ɸaɰai* 'bandicoot', and *sije* 'pig' from recorded elicitation items in 20111208-6-kiy-ap-nps-vps; Figure 7a (top) shows the contours for all the words in a single plot, while Figure 7b (bottom) partitions the $f_0$ contours by target word into separate panels.

The lack of structure in the set of $f_0$ contours in Figure 7a in contrast to the orderliness of the set of $f_0$ contours plotted by target word in Figure 7b suggests that properties of the target word are a large source of variability in the pitch contour. However, the substitution frame and associated properties are not a large source of variability, since $f_0$ contours elicited in different substitution frames for the same target word are very similar. A glance at the marked difference in Figure 7b between $f_0$ contours for *naβi* and *ɸaɰai*, despite their sharing the same LENGTH and LEXICAL CLASS, is strong evidence that we are missing some property of the target word, some latent variable, that is a large source of variability in the pitch contour: namely, TONAL CLASS.

**2.3 EXPERIMENTAL DESIGN IN LATER TONEME DISCOVERY: TONEMES AS INDEPENDENT VARIABLES.** In §2.2, the *explanatory variables* are a set of variables (in Table 1) that we think could influence pitch contours over target words within an elicitation item. This is because the hypothesis in play at that point, Hypothesis 3 (repeated below), is that those variables can explain the variability in pitch contours over words:

Hypothesis 3: (Explaining variation in pitch contours over words) *The primary influences on variation in pitch contours over words are the variables listed in Table 1.*

Once we have enough experience with tonal classes to group words by their putative tonal classes, we can begin to treat (putative) TONAL CLASS (of the target word, i.e. Pike's substitution item) as an *independent variable* to be systematically manipulated, alongside the set of independent variables given in Table 1. Our research question and hypothesis at this stage of toneme discovery become:

Question 2: (The effect of tonal class on pitch contours) *Does TONAL CLASS affect the pitch contour of a word?*

Hypothesis 4: (Explaining variation in pitch contours over words) *A primary influence on variation in pitch contours over words is TONAL CLASS.*

Asking this research question introduces a new partition in our set of independent variables: the partition between *explanatory* and *confounding* variables (also called *extraneous* or *nuisance* variables). Since what we are interested in is the effect of TONAL CLASS on the pitch contour, TONAL CLASS is an explanatory variable. All the other independent variables (those enumerated in Table 1) which were treated as explanatory variables in §2.2 are now treated as confounding variables with respect to Question 2. (See Snider, this volume, for further discussion of confounding variables in establishing underlying tonal contrast.) In the experimental design when TONAL CLASS was not yet on the table as an independent variable (§2.2), all the independent variables were treated as being explanatory variables to see if we could explain away all the variability in pitch contours over words with them (Hypothesis 3), without resorting to the latent variable of TONAL CLASS. But now with TONAL CLASS as the variable of primary interest, these other variables could obscure the relation between pitch contour variability and TONAL CLASS. Thus, they are confounding.

What is an explanatory variable and what is a confounding variable always depends on the research question. To abuse the old saw about junk and treasure, "one experiment's confounding variable is another experiment's explanatory variable." Both explanatory and confounding variables are systematically manipulated in an elicitation experiment, however, so both types of variables—even the confounding variables—are sources of explained variability in the decomposition of variability illustrated in Figure 4.

In this new experimental design, we treat TONAL CLASS differently from the confounding variables, as it is the sole explanatory variable, and compare the proportion of variability explained with TONAL CLASS included as an independent variable to the proportion explained when it is not included. We deal with the confounding variables with strategies following the classic work of Fisher (1925, 1935) which laid the foundations of experimental design. Fisher proposed three strategies for reining in the effect of confounding variables on the dependent variable: blocking, replication, and randomization. With these strategies, the hypothesis that "if two words have different pitch contours, then they belong to different lexical tonal classes" (Hypothesis 2) becomes reasonable rather than ridiculous. We can update Hypothesis 2 as follows:

Hypothesis 2': Linking hypothesis between lexical tonal classes and pitch contours

Lexical tonal classes (tonemes) induce systematic variation in the pitch contours of words. Therefore, the unobservable concept of a toneme is observable via its influence on the pitch contours of words: *provided that other confounding sources of variability in the pitch contour are taken into account,* if two words have sufficiently different pitch contours, then they belong to different lexical tonal classes.

The notion of *sufficiently* different pitch contours takes into account that even within a controlled context, pitch contours within a single tonal class are distributed with some variability (see any of the small boxes for Tone 2 and 3 in Figure 8)—variability that is unexplained by the variables considered in the elicitation design. We have reason to assume that this within-class variability is small relative to the variability between tonal classes (Deutsch et al. 2004), especially within a controlled context within an elicitation session and within a speaker. Geometrically, the assumption is that the distribution of pitch contours for a given tonal class is tightly clustered, while the distance between clusters for different tonal classes is much larger than the distance spanned within a cluster. Ultimately, whether or not pitch contours are sufficiently different for tonemic contrast must be perceptually confirmed with the consultant.

**2.3.1 BLOCKING AND REPLICATION.** Pike's strategy of dividing elicitation items into uniform groups is an example of *blocking* confounding variables and was conceived as such:

> In order to be significant the tonal contrasts must be found in words which are sufficiently similar to rule out interference from nonpitch characteristics, and they must occur in contexts which cannot cause the observed pitch differences. (Pike, 1948, p. 48)

For instance, elicitation items are split into homogeneous groups by the confounding variable LENGTH of the word: one block consists of only monosyllables, another block of only disyllables, etc. Within each of these blocks, the level of LENGTH is held constant, and each elicitation item is assigned to one level of the explanatory variable TONAL CLASS, e.g. Tone 1. Within an elicitation session, elicitation items are organized by these blocks: all items within one block are elicited before moving onto the next block.

All the variables in Table 1 are treated by Pike as blocked confounding variables, i.e. blocking variables. Although there are multiple blocking variables, we can create a single aggregate blocking variable called BLOCK subsuming all the confounding variables in Table 1.[14] Thus, a block consists of a homogeneous group of elicitation items, matched for every variable listed in Table 1, e.g. one level of the BLOCK variable might be the group of items specified by the fixed levels in Table 4. This is like a generalized substitution frame, a complete fixed aggregate specification of context for an elicitation item.

Let's expand on this notion of a generalized substitution frame: while Pike refers to only the substitution frame as "context" in the quote about blocking above, I extend "context" here to refer to the morphological and phonological properties of the words from the first step in the procedure as well: since we would like to assume that any variation in pitch contours between words is due solely to tonemic contrast, *any* aspect of the elicitation situation that affects the pitch of a word, but which isn't tonal class, is part of the "context". We abstract away from these contextual factors in two ways: (1) by explicitly regulating some as confounding variables via strategies such as blocking to contribute to explained variability, and (2) by leaving some unregulated to affect the pitch contour as they will, thus contributing to unexplained variability.

| Variable | Fixed levels |
|---|---|
| Lexical class | Noun |
| Morphological complexity | Simplex |
| Length | Disyllabic |
| Syllable shape | CVCV |
| (Vowel length) | (short) |
| Segmental features | initial voiceless obstruent |
| Prosodic position | utterance-initial |
| Pragmatic context | Out-of-the-blue focus |
| Syntactic structure | Possessive prenominal phrase |

TABLE 4: An example of a block specification aggregated over the confounding variables in Table 1. Each of the confounding variables is fixed at the levels stated in the table, e.g. LEXICAL CLASS in the block is fixed to be "noun". We might call this particular specification of the confounding variables *Block 1*, one of multiple levels of the aggregate blocking variable BLOCK. Another block, say *Block 2*, might have identical specifications other than that LEXICAL CLASS fixed as "verb" rather than "noun".

[14] There are more sophisticated ways to incorporate multiple blocking variables into an experimental design, such as the Latin square designs commonly used in psycholinguistics (Maxwell and Delaney 2004, p. 557-561, Montgomery 2005, p. 136–142), but it's sufficient for our conceptual discussion here to refer to a single aggregate blocking variable.

By blocking, we can take advantage of our knowledge of some of the sources of variability in the pitch contour of a word. Rather than leaving the variability induced by those sources as unexplained, we parcel that variability out between blocks, i.e., as variability explained by the blocking variable, thus adding to the proportion of variability explained. Previously, when we treated TONAL CLASS as a latent variable (§2.2), explained variability consisted only of variability explained by variables aggregated in BLOCK. Now, the explained variability consists of variability explained by the explanatory variable TONAL CLASS as well as variability explained by the confounding variables aggregated in BLOCK.

Blocking reduces the noise in the observed effect of TONAL CLASS on pitch contours. Across blocks, the effects of TONAL CLASS on pitch contours might look quite different, so that the particular pitch contours induced over words by different tonal classes may be quite different. But within a homogeneous block, systematic variability in pitch contour induced by TONAL CLASS is much clearer since variability in pitch contours within a block due to factors other than TONAL CLASS is small. What is uniform about the effect of TONAL CLASS *across* blocks is not the particular pitch contours induced by a given tonal class, but the unified behavior of members of a tonal class within each block.

An example of how blocking can reduce unexplained variability in the pitch contour is schematized below in Figure 8, which shows pitch contours for two putative tonal classes in Kirikiri, Tone 2 and Tone 3, realized in two different blocks (see §2.3.3 for more details on the data). Within Block 1, the variability in pitch contours due to factors other than TONAL CLASS is small: for instance, whether or not the class is Tone 2 or Tone 3 fully determines whether the pitch contour in Block 2 is a steep rise or a shallow fall. However, between blocks, the variability in the pitch contour within a single tonal class is large: for instance, the pitch contour for Tone 2 is a fall-rise shape in Block 1, but a steep rise in Block 2. The role of TONAL CLASS in inducing systematic variability in pitch contours would be obscured without partitioning the data into blocks.

An additional benefit of blocking is that each block serves as a *replication* of the experiment to confirm the way that TONAL CLASS affects pitch contours of target words. Converging evidence from each replication about systematic variability in pitch contours due to TONAL CLASS boosts our confidence about our posited tonal contrasts.

However, sometimes fixing a confounding variable at a constant level for the whole (elicitation) experiment, e.g. running just a single block in an experiment with one confounding variable, is also a good option. One can also run a subset of multiple levels among all the possible levels, but this is less common than running either all of the levels or just one, since it can be hard to justify why one picked some levels but not others.

As an instance where fixing a variable at a constant level might be desirable, consider the variable SONORANCY (of all consonantal segments in the target word), with the two levels [+sonorant], [−sonorant]. One could vary between these two levels between blocks, but one could also fix SONORANCY to be [+sonorant]. This is very common in tonal and intonational studies, since working exclusively with [+sonorant] segments can help reduce segmental perturbations to the $f_0$ and pitch contour when tone-segment interactions are not of interest. Abstracting away from [−sonorant] consonants has the disadvantages that we miss the opportunity to: (1) replicate our elicitation experiment by varying sonority to build our confidence in our conclusions about toneme discovery, and (2) study interactions between levels of SONORANCY and TONAL CLASS. But by fixing SONORANCY at

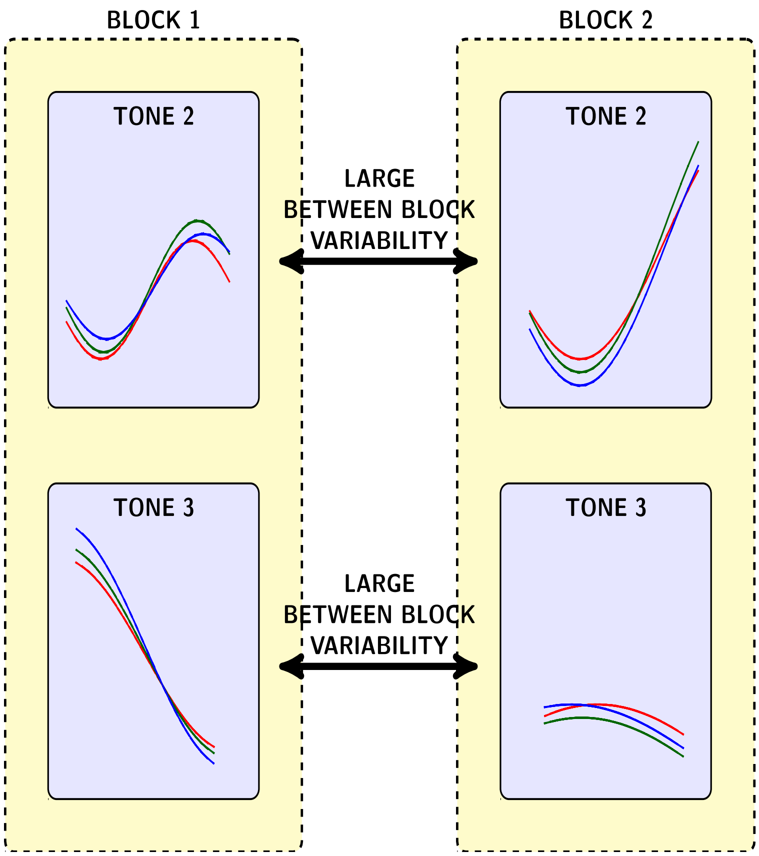

BLOCK 1
BLOCK 2
TONE 2
TONE 2
LARGE BETWEEN BLOCK VARIABILITY
TONE 3
TONE 3
LARGE BETWEEN BLOCK VARIABILITY

FIGURE 8: Schematic illustration of how blocking can reduce unexplained variability in the pitch contour, adapted from the Kirikiri data set in §2.3.3. Between blocks, the pitch contours for Tones 2 and 3 are quite different, but within blocks, the pitch contours for each tonal class are quite homogeneous. Blocking thus helps reveal the systematic variability induced in pitch contours by TONAL CLASS.

[+sonorant], we have the advantages of: (1) abstracting away from noise introduced by segmental perturbation in the relation between TONAL CLASS and the pitch contour that is not of interest for the current research question, and (2) reducing the number of items in the elicitation so that the elicitation isn't too long and grueling.

Whether it's best to hold a confounding variable fixed or to vary it depends entirely on the research question. One rule of thumb for factors to consider in the decision, if there are no other compelling reasons based on the research question, is to consider either choosing the "vanilla" level or choosing two or more extreme levels.[15] By "vanilla" level, we mean, roughly speaking, the most unmarked level. For instance, for SYLLABLE SHAPE, this might be

[15] I am grateful to Pat Keating for teaching me this.

CV rather than CCV, CV:, or CVC, depending on the word-prosody of a language. Picking a “vanilla” level can be a good option when we are wholly uninterested in the variable and simply need to choose something in order to carry out the elicitation experiment. Picking at least two extreme opposing levels, e.g. liquid and voiceless plosive for ONSET SONORITY (with sonority conceived of as a gradient scale), can be a good option when we are more interested in the interaction between the explanatory variable(s) and the confounding variable at hand—perhaps we’re worried about generalizing our conclusions across the levels of the confounding variable, so performing replications at one or more extreme levels provides converging evidence for our conclusions and/or a challenging test of our hypotheses under worst case limiting conditions.[16]

A final important clarification about blocking in elicitation design is the distinction between blocking for speaking by the consultant and blocking for listening by the fieldworker. For Pike, these two aspects of elicitation were conflated since his discussion of toneme discovery largely assumed that the opportunity to listen to tones only comes with live elicitation. However, the ready availability of recording technology in the present day highlights that there is indeed a distinction between the elicitation design for the consultant’s speaking and the fieldworker’s listening. (The tutorial at http://www.krisyu.org/blog/posts/2013/06/recording-in-the-field/ gives some tips on recording practices in the field, and the two tutorials at http://www.krisyu.org/blog/posts/2013/06/processing-audio-files-praat/ and http://www.krisyu.org/blog/posts/2013/06/processing-audio-files-sox/ provide information on preparing recorded audio files for analysis.) For instance, given an annotated recording of an elicitation session, the software package Toney (See Bird and Lee, this volume) provides an easy way to perform post-hoc blocking *for the purpose of listening to the elicited items*, i.e. blocking of items from an elicitation session after the session is completed. (The tutorial at http://www.krisyu.org/blog/posts/2013/06/annotating-audio-files/ provides an introduction to annotating audio files.) Even if blocking was not imposed during the consultant’s productions in the elicitation session, Toney can extract clips of items and play them such that one can hear all the items in a post-hoc block one after the other.

In toneme discovery, blocking in listening to elicited items is essential for allowing the systematic variability induced by TONAL CLASS to emerge from the swamp of variability induced by other factors, as we’ve discussed in this section. But how about the effect of blocking on the consultant’s productions? In some cases, the homogeneity introduced by blocking can be helpful for the consultant. One general such case is when it may be difficult for the consultant to get into the frame of mind desired by the fieldworker for a particular elicitation context. For instance, suppose that the set of items for elicitation consists of deeply embedded possessive constructions like ‘monkey’s enemy’s snake’s ear’ (Hyman 2007). When asked to produce ‘monkey’s enemy’s snake’s ear’ out-of-the-blue with English as the contact language, the consultant might interpret the phrase as the simpler and unintended conjoined phrase ‘monkeys, enemies, snakes, ear’. The intended possessive constructions may be made understood to the consultant with a story behind each individual item, but for the sake of efficiency in the face of time limitations, the fieldworker may want to introduce the construction with a more familiar possessive example such as

[16] The strategy of testing extreme cases is common for error checking and debugging in mathematics and programming, too.

'John's father's friend's dog', and use this particular item to *prime* the consultant to keep producing the intended possessive parse for all items within the block. Thus, priming a particular parse within a block may be a useful design tool in elicitation.

But priming can also be undesirable in cases where it induces effects in production that are an unintended artifact of the priming. For instance, if a consultant is asked to produce a long list of trisyllabic words, all of which are known to be Tone 2 words, followed by some words elicited for the first time, it's possible that the consultant might unintentionally produce these new words as Tone 2 words, simply because they were preceded by many Tone 2 words. This would result in the fieldworker's mistaken classification of the new words as Tone 2 words based on the elicitation data. Another elicitation design feature in addition to blocking that can be useful to avoid such unintended priming effects is *randomization*. This may sometimes also be the only choice for dealing with confounding variables when blocking is not possible.

**2.3.2 RANDOMIZATION.** Some confounding variables are not amenable to blocking, so instead of blocking by such variables, we randomize over such variables within a block. Consider the order of elicitation items:

> The mere fact that one word is necessarily said before the other in repetitions by the informant will frequently cause sandhi changes, or phrasal conditioning, or intonational modifications of one of the words. To check on this possibility the investigator should (1) reverse the order in which the items are repeated, and (2) have the informant make a marked pause before each item. (Pike, 1948, p. 54)

As Pike states, it is unavoidable that the elicitation context sets up a discourse context in which the discourse extends beyond single elicitation items. Thus, prosodic marking of prominence and demarcation (phrasing) due to the imposition of prosodic structure and/or the particular pragmatic context introduced in the current discourse context may induce variation in pitch contours. A particularly dire case is if there is a bias in where elicitation items of a given tonal class appear in the order of items. Suppose that the tonemes of a tone language include a falling tone as well as a low tone. Now suppose that within each block, we have one item exemplifying each toneme, and that the low tone items are always elicited at the end of the block. Since low tones often fall utterance-finally due to the interaction of tonal and intonational effects on the pitch contour (Maddieson 1978, p. 45-46), we might not be able to distinguish the true falling tone and the low tone.

What would be the levels of an independent variable ORDER (of tonal class)? Suppose we had five tonal classes, and one exemplar of each tonal class to use as a substitution item within a block such as the block specified in Table 4. With only five items, we would have 5! = 5 × 4 × 3 × 2 × 1 = 120 possible orders of the items within the block! Adding elicitation order as a blocking variable would add 120 times as many blocks. Another problem with blocking by ORDER is that it's not clear that this is the way we would want to define elicitation order—perhaps all that matters is what TONAL CLASS level is last in a block, and we could randomize the order of all other stimuli in the block. Then we would only have five levels of ORDER, one level for each of the five elicitation items occupying the last slot in the block. Or alternatively, perhaps what matters is the identity of the TONAL CLASS of two elicitation items that occur next to one another in the elicitation sequence.

An alternative to blocking is *randomization*:

> Randomisation properly carried out … relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed. (Fisher, 1935, p. 44)

Rather than blocking by ORDER, rather than attempting to make sure that there are no biases in ordering with respect to TONAL CLASS by painstakingly fiddling with orders by hand, we can *randomize* order of elicitation items within a block to eliminate bias. One way to conceptualize this is the following: for each block, suppose we label each elicitation item within that block with an integer, e.g. 1, 2, 3, …, 24, 25 if there are 25 items in the block. We then write that number on a slip of paper and put all of the slips in a jar and mix them up. Each time we're ready to elicit a new item within a block, we draw a slip from the jar and read the number written on it. The number on the slip tells us which item to elicit. We then put the slip in the recycling bin (not back in the jar) and elicit the corresponding item.

With the introduction of randomization into our experimental toolbox, we return to the hypothetical example from §2.3.1 where priming from producing a long list of trisyllabic Tone 2 words resulted in the unintentional production of another word as a Tone 2 word. One way to help prevent the unintentional priming effect would be to keep the Tone 2 block, but to also include *fillers* as elicitation items in the block and to *randomize* the order of all the items, including the fillers, within the block. The fillers would be items distracting the consultant from the common properties of the block, e.g. items from other tonal classes, but still sharing common properties with the Tone 2 words of interest, e.g. also trisyllabic. The fillers might even be items of interest for another research question, echoing the idea that "one experiment's confounding variable is another experiment's explanatory variable." Rather than perform a simple randomization of all the items, we might also consider *pseudo-randomization* (constrained randomization), and generate a randomized order subject to the constraint that the first and last items in the list are fillers. This is often done to keep effects due to initiating and ending a list from affecting the items of interest.

This concludes the discussion of the second stage in toneme discovery, where (putative) TONAL CLASS is treated as an independent variable. In this section, we saw how to use strategies of blocking, replication, and randomization to keep confounding variables from obscuring the relation between TONAL CLASS and the pitch contour over a word. In §2.3.3 below, we illustrate an example of manipulating TONAL CLASS as an independent variable in Kirikiri, focusing on blocking and replication as strategies. The tutorial on preparing items for elicitation at http://www.krisyu.org/blog/posts/2013/06/preparing-elicitation-items-for-presentation/ shows how to perform simple randomizations of elicitation order with examples from Kirikiri elicitations with a spreadsheet. There are also web-based tools for simple randomization available, such as http://www.random.org and http://www.randomizer.org/.

#### 2.3.3 EXAMPLE: TONAL CLASS AS AN INDEPENDENT VARIABLE IN KIRIKIRI.

Section 2.2.1 gave an example of discovering TONAL CLASS as a latent source of variability in the pitch contour in Kirikiri. In this section, I fast forward to one of the first elicitation sessions where the fieldwork team explicitly treated TONAL CLASS as an independent variable and systematically ran exemplars of different proposed tonal classes through a series of substitution frames in $N_1 + N_2$ prenominal possessive constructions ($N_1$*'s* $N_2$), e.g. *ɸaɰai giɰʉ.* 'bandicoot's elbow' composed from *ɸaɰai* 'bandicoot' and *giɰʉ* 'elbow'. This is elicitation session 20111213-kiy-ap-1-framedwordlist, and a full list of elicitation items from the session is given in Table A.2 in the Appendix.

The substitution frames were chosen to treat TONAL CLASS of the substitution frame, a single word, as a blocking variable. This is so that each of the five proposed tonal classes for the substitution frame serve as a block for replication for testing the proposed tonal classes of the substitution items (target words). Thus, while TONAL CLASS *for the substitution item* is treated as an explanatory variable, TONAL CLASS *for the substitution frame* is treated as a confounding variable. In fact, the experimental design has two experiments in one: we can consider either $N_1$ to be the substitution item and $N_1$ TONAL CLASS as the explanatory variable and $N_2$ to be the substitution frame and $N_2$ TONAL CLASS a confounding variable, or vice versa. (See §2.4 for an alternative conception of this experimental design.) When the target word is $N_1$, it's utterance-initial, and when it is $N_2$, it is utterance-final, so we have experimental replications over different prosodic positions, as well as replications over different flanking tonal classes. The experimental design for this session, taking $N_1$ to be the substitution item, is given below. The experimental design taking $N_2$ to be the substitution item is nearly identical: wherever $N_1$ occurs in the description of the experimental design, replace it with $N_2$.

- Research question: How does TONAL CLASS affect pitch contour over a word?
- Strategy: Manipulate TONAL CLASS as an independent variable in different substitution frames.
- Research hypothesis: There are five tonal classes in Kirikiri.
- Linking hypothesis: We assume that distinct pitch contours imply distinct levels of TONAL CLASS, as in Hypothesis 2. Moreover, we assume that in one or more substitution frames, the pitch contour for a proposed tonal class is distinct from pitch contours for other proposed tonal classes.
- Experimental unit: individual elicitation items
- Explanatory variables: TONAL CLASS (of target word $N_1$), with levels *T1, T2, T3, T4, T5*
- Confounding variables
- WORD LENGTH (of target word): 2 syllables (fixed at this level)
- SYNTACTIC STRUCTURE: prenominal possessive phrases (fixed)
- TONAL CLASS OF FRAME WORD $N_2$, with levels *T1, T2, T3, T4, T5* … (blocking variable)
- PROSODIC POSITION (of substitution item): utterance-initial (fixed; utterance-final if $N_2$ is the target word)
- WORD LENGTH (of substitution frame): 2 syllables (fixed)
- Dependent variable: $f_0$ contour over the target word $N_1$

A list of the five different blocks for each of the two sub-experiments, one with $N_1$ as the target word, and one with $N_2$ as the target word, is given in Table 5. The full layout of tonal sequences in the experimental design is given in Table 6. The levels of the independent variables $N_1$ TONE and $N_2$ TONE are fully cross-classified. We'll see another way to think about these kinds of fully cross-classified designs in §2.4. The exemplar words chosen for each TONAL CLASS for $N_1$ and $N_2$ are given in Table 7.

| Block | Independent variable | Levels |
|---|---|---|
| $N_2$: T1 | $N_1$ TONE | T1, T2, T3, T4, T5 |
| $N_2$: T2 | $N_1$ TONE | T1, T2, T3, T4, T5 |
| $N_2$: T3 | $N_1$ TONE | T1, T2, T3, T4, T5 |
| $N_2$: T4 | $N_1$ TONE | T1, T2, T3, T4, T5 |
| $N_2$: T5 | $N_1$ TONE | T1, T2, T3, T4, T5 |
| $N_1$: T1 | $N_2$ TONE | T1, T2, T3, T4, T5 |
| $N_1$: T2 | $N_2$ TONE | T1, T2, T3, T4, T5 |
| $N_1$: T3 | $N_2$ TONE | T1, T2, T3, T4, T5 |
| $N_1$: T4 | $N_2$ TONE | T1, T2, T3, T4, T5 |
| $N_1$: T5 | $N_2$ TONE | T1, T2, T3, T4, T5 |

TABLE 5: Manipulated independent variables in the experimental design for elicitation session 20111213-1-kiy-ap-framedwordlist. TONAL CLASS for $N_1$ ($N_2$), the target word, is varied over the 5 putative tonal classes within each block of varying the tonal class of the substitution frame word, $N_2$ ($N_1$).

| | *+T1* | *+T2* | *+T3* | *+T4* | *+T5* |
|---|---|---|---|---|---|
| T1 | T1 + T1 | T1 + T2 | T1 + T3 | T1 + T4 | T1 + T5 |
| T2 | T2 + T1 | T2 + T2 | T2 + T3 | T2 + T4 | T2 + T5 |
| T3 | T3 + T1 | T3 + T2 | T3 + T3 | T3 + T4 | T3 + T5 |
| T4 | T4 + T1 | T4 + T2 | T4 + T3 | T4 + T4 | T4 + T5 |
| T5 | T5 + T1 | T5 + T2 | T5 + T3 | T5 + T4 | T5 + T5 |

TABLE 6: Manipulation of TONAL CLASS as an independent variable in a sequence of two tones. The levels of the independent variables $N_1$ TONE (across rows) and $N_2$ TONE (across columns) are cross-classified so that all possible combinations of tones (5×5=25 in total) are included.

| Tone | Noun | Word |
|---|---|---|
| T1 | N1 | ɸaɰai ‘bandicoot’ |
| | N2 | giɰʉ ‘elbow’ |
| T2 | N1 | kaza ‘gecko’ |
| | N2 | oɾa ‘tongue’ |
| T3 | N1 | ɸiʋa ‘snail’ |
| | N2 | ɸaɰao ‘groin’ |
| T4 | N1 | nala ‘wasp’ |
| | N2 | kʍawaa ‘chin’ |
| T5 | N1 | taʋa ‘catfish’ |
| | N2 | kʊɾee ‘string’ |

Table 7: Wordbank for $N_1 + N_2$ data set from 20111213-1-kiy-ap-framedwordlist.

In Figure 9 below, we show $f_0$ contours for all words, whether they were from the first noun ($N_1$) or the second ($N_2$). This figure gives a baseline snapshot of the total variability present in the $f_0$ contours and is unsurprisingly chaotic like Figures 5a and 6a. All $f_0$ contours plotted in this section are *time-normalized*. (See the materials used to prepare the data for this paper at http://www.krisyu.org/blog/posts/2013/10/preparing-data-for-ldc-paper/ for more information. One useful software package for producing time-normalized $f_0$ contours using Praat is Yi Xu’s ProsodyPro, available at http://www.phon.ucl.ac.uk/home/yi/ProsodyPro/.) For each $f_0$ contour, mean $f_0$ was extracted from 30 evenly spaced frames or samples over the word. The effect of this normalization is to improve comparability of $f_0$ contours over words that may have different durations and/or have been uttered at different speech rates.

Time-normalized $f_0$ contours plotted separately by putative tonal class but irrespective of block are shown in Figure 10 below. Whether the $f_0$ contour comes from the first or second noun, $N_1$ or $N_2$, is indicated by line color and point shape. Plotting $f_0$ contours in individual panels for each putative tonal class helps reveal the systematic effect of tonal class on variability in $f_0$ contours: for instance, T4 $f_0$ contours all share a common rise-fall shape, distinct from the more level $f_0$ contours of T1. But within each panel, $f_0$ contours for the same tonal class may still be divergent. Most of this divergence can be traced to whether the $f_0$ contour came from $N_1$ or $N_2$. For instance, T2 in $N_1$ shows a large fall at the end of the word, but the same tone in $N_2$ has a large pitch rise to the end of the word, and T3 in $N_1$ has a large early peak, while T3 in $N_2$ has no large peak. Is the divergence between the $f_0$ contours from $N_1$ and $N_2$ traceable to an effect of prosodic position, i.e. whether the word is utterance-final (as in $N_2$) or not? Figure 10 shows that T1, T3, and T4 all appear to be subjected to an overall lowering of $f_0$ in $N_2$ relative to in $N_1$, which would be consistent with effects expected by utterance-final lowering and downtrends over a prosodic phrase. However, T2 and T5 show $f_0$ contours that are higher rather than lower at the end of word for $N_1$ than $N_2$, behavior not consistent with this prosodic explanation, unless it is the case that prosodic position-specific lowering may be conditioned by tone.

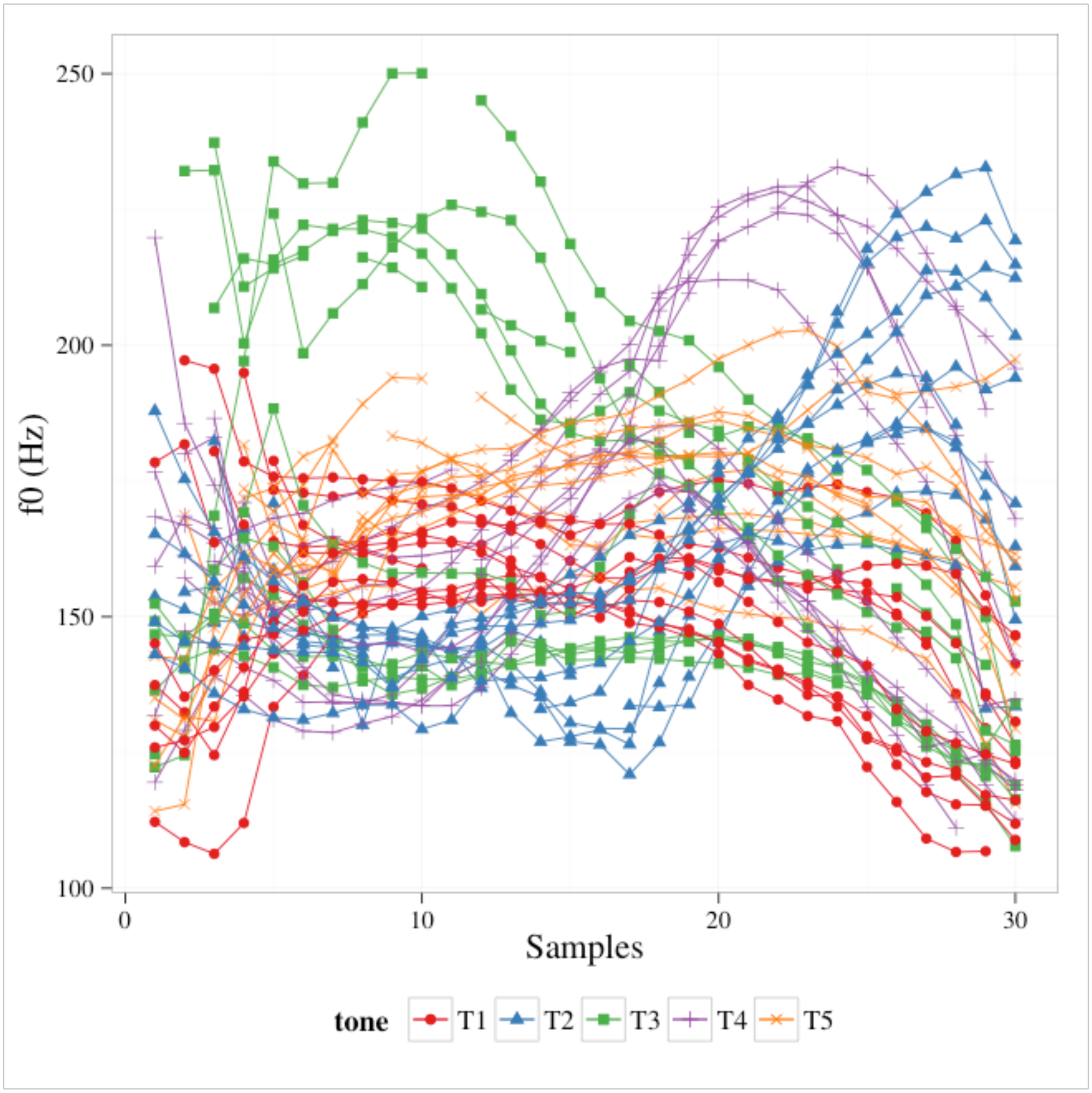


FIGURE 9: A plot of time-normalized $f_0$ contours from all target words elicited in the session 20111213-1-kiy-ap-framedwordlist for both $N_1$ and $N_2$. Tonal class is indicated by color and point shape. The x-axis indicates *samples*, not an absolute time scale, since the $f_0$ contours are time-normalized. For each word, 30 evenly spaced samples were taken from the $f_0$ contour.

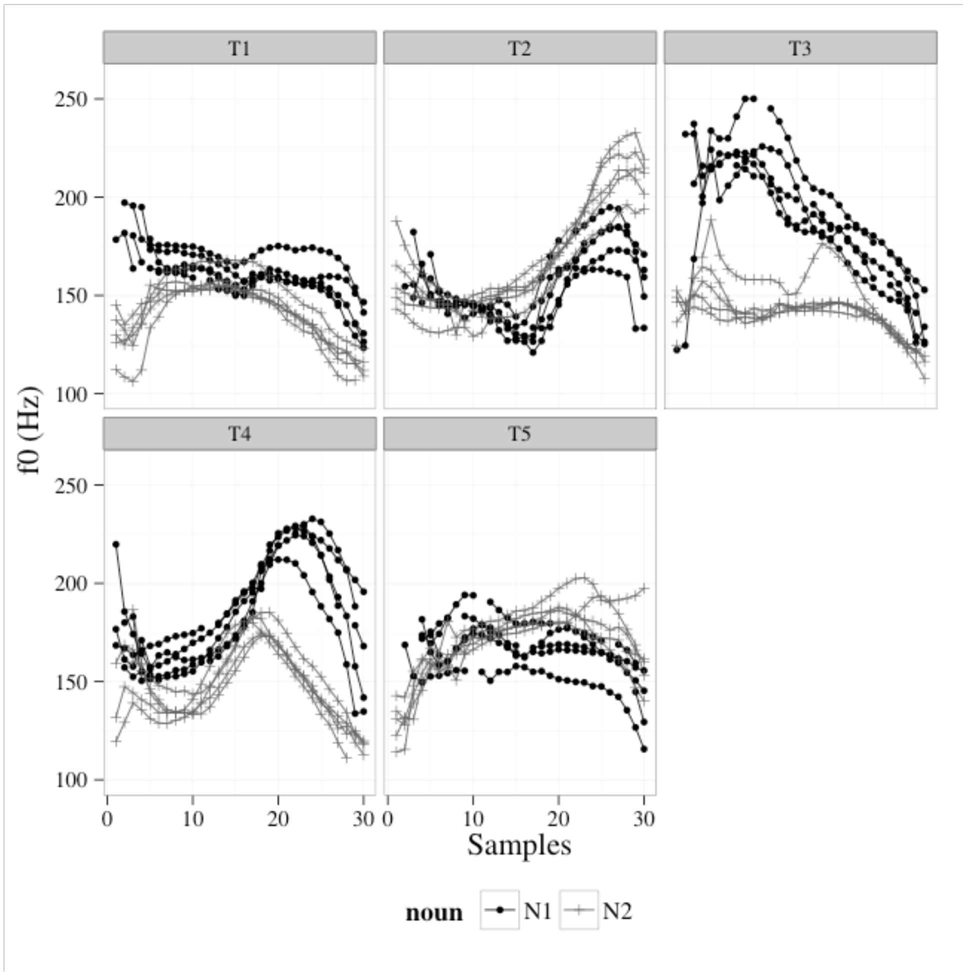


Figure 10: Time-normalized $f_0$ contours over target words plotted separately for each putative tonal class. Words pronounced utterance-initially ($N_1$) are indicated with black lines and points, while words pronounced utterance-finally ($N_2$) are indicated with gray lines and crosses.

Figure 11 suggests an alternative explanation. This figure shows the $f_0$ contour across the entire $N_1 + N_2$ construction, broken into two panels. Each row of the figure shows all the $f_0$ contours for a given, fixed $N_2$ tonal class indicated on the label on the right side of the right panel; the left panels display the $f_0$ contour over $N_1$, and the right panels display the $f_0$ contour over $N_2$. The left panels in Figure 11 show that regardless of what the $N_1$ tonal class is, the $f_0$ contour over $N_1$ tends to end heading downwards towards a low $f_0$, and the right panels show that $f_0$ at the onset of $N_2$ is also uniformly low across $N_2$ tonal class. Thus, there appears to be a low tone or at least a low tonal target at the boundary between $N_1$ and $N_2$.

Regardless of the exact shape of the $f_0$ contour for a given tonal class and how it may be conditioned in different environments, the example of elicitation design and analysis explicated in this section demonstrates how blocking can help reveal systematic effects of the explanatory variable, i.e. TONAL CLASS (of the target word) on variability in $f_0$ contours. This is most clearly visible in Figure 10, which shows large between-block variability but small within-block variability for $f_0$ contours within a tonal class, if the blocking is taken to be whether the $f_0$ contour is from the first noun, $N_1$, or the second, $N_2$. This partition of blocks is coarser than the one originally intended in Table 5, which was a partition by tonal class of the flanking noun. Perhaps the particular tonal class of the flanking noun doesn't have significant effects on the $f_0$ contour in the elicited possessive constructions because of the presence of an unexpected low tonal target between the nouns. Follow-up work would need to be done to clarify the issue, e.g. to check that the low tonal target isn't due to segmental perturbation. This example is thus characteristic in demonstrating the iterative nature of hypothesis generation and testing in elicitation.

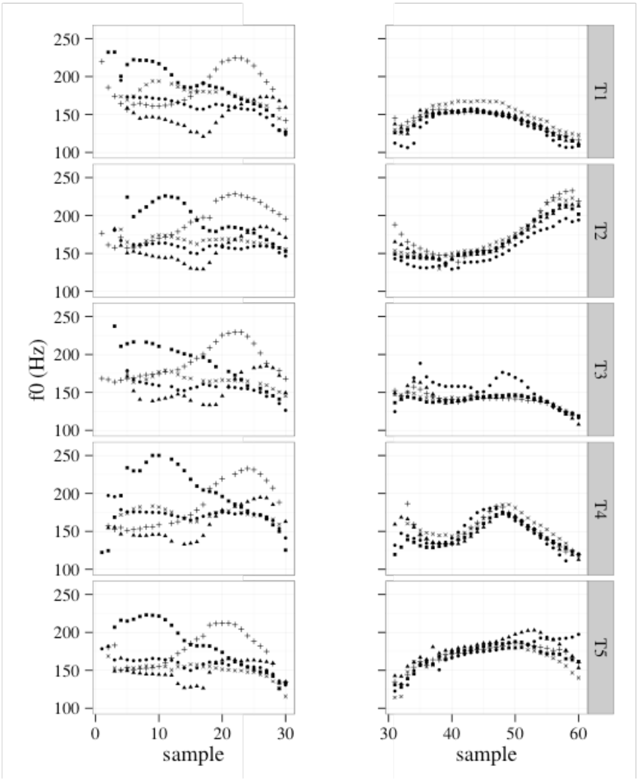


FIGURE 11: Time-normalized $f_0$ contours over the entire $N_1 + N_2$ construction, split into left panels displaying the $f_0$ contour over $N_1$ and right panels displaying the $f_0$ contour over $N_2$. Each row shows all $f_0$ contours for a fixed $N_2$ tonal class indicated on the label on the right panel. $f_0$ contours tend to head towards a low $f_0$ between $N_1$ and $N_2$, regardless of $N_1$ and $N_2$ tonal classes.

**2.4 USING EXPERIMENTAL DESIGN PRINCIPLES TO GENERALIZE BEYOND PIKE (1984): TONOTACTICS IN THLANTLANG LAI (HYMAN, 2007).** In §2.2 and §2.3, we saw how Pike's toneme discovery procedure can be recast in terms of general principles of experimental design. Understanding these principles allows us to connect Pike's elicitation methodology with other methods in elicitation design and to generalize Pike's methodology for very different research questions. As an example of this, we illustrate how Hyman (2007)'s work on tonotactics in Thlantlang Lai (Tibeto-Burman Kuki-Chin, Myanmar) uses the strategy of inferring the presence of hidden structure by treating tonotactic constraints as a latent variable, as in §2.2, and how it uses the same experimental design as the example illustrating toneme discovery in Kirikiri prenominal possessive phrases in §2.3.3. For other examples of designs, see http://www.krisyu.org/blog/posts/2013/10/additional-examples-of-elicitation-designs/.

In tonal fieldwork, as soon as the fieldworker is far along enough in toneme discovery to manipulate TONAL CLASS as an explanatory independent variable, a natural set of follow-up research questions pursues further detail in understanding tonal allophony and alternation. Here, we take tonal allophony as an example. Hyman (2007) examined the surface tone sequences of prenominal possessive phrases, e.g. *râal ràng* 'enemy's horse'.

We can cast this in terms of the following experiment:

- Research question: Are there tonotactic *n*-gram restrictions for surface tones in Thlantlang Lai? [17]
- Strategy: Control any variables suspected to induce variation in surface realization of underlying tones. (Tonotactic restrictions, perhaps formulated as constraints, are not included in this set of variables, since whether or not Thlantlang Lai has such restrictions is yet unknown).
- Research hypothesis: There are tonotactic bigram restrictions in Thlantlang Lai.
- Linking hypothesis: We assume some mapping between the pitch contour over the word and the surface tones.
- Experimental unit: *n*-gram over Thlantlang underlying tones, e.g. $N_1 + N_2$ bigram
- Explanatory variables: UNDERLYING TONAL CLASS of each noun: $N_1$ TONE, $N_2$ TONE, …, $N_n$ TONE, with levels $H, L, HL$
- Confounding variables
- *n*-GRAM LENGTH: bigrams, trigrams, 4-grams, … (blocking variable)
- SYNTACTIC STRUCTURE: prenominal possessive phrases (fixed at this level)
- WORD LENGTH (syllables): 1 syllable (fixed at this level)
- PROSODIC POSITION (of *n*-gram): isolation (fixed)
- Dependent variable: surface tone sequence of the *n*-gram

This strategy should seem familiar: just as in §2.2, we try to control for everything we suspect may affect the dependent variable and then if we still have significant residual variability left unexplained, we attribute that to hidden structure—in this case, tonotactic

[17] An *n-gram* is a sequence of discrete units of length *n*, e.g. a 2-*gram* or *bigram* is a sequence of length 2.

constraints. Thus, we have a number of blocking variables, most which we fix to a single level. (For example, SYNTACTIC STRUCTURE is held fixed as prenominal possessive phrase.

Like in Pike's experimental design in §2.3, an explanatory variable is (underlying) TONAL CLASS. However, here we have multiple explanatory variables, since we are interested in sequences of tones—we have one explanatory variable per noun, e.g. for bigrams, we have two explanatory variables, each with three levels: HL, H, and L. Table 11 shows the explanatory variables for each *n*-gram block and Table 12 shows the words used to create the bigram sequences. Each *n*-gram block provides a replication for discovering bigram tonotactic restrictions.

| Block | Explanatory Variable | Levels |
|---|---|---|
| $N_1 + N_2$ | $N_1$ TONE | HL, H, L |
| | $N_2$ TONE | HL, H, L |
| $N_1 + N_2 + N_3$ | $N_1$ TONE | HL, H, L |
| | $N_2$ TONE | HL, H, L |
| | $N_3$ TONE | HL, H, L |
| $N_1 + N_2 + N_3 + N_4$ | $N_1$ TONE | HL, H, L |
| | $N_2$ TONE | HL, H, L |
| | $N_3$ TONE | HL, H, L |
| | $N_4$ TONE | HL, H, L |
| ⋮ | ⋮ | ⋮ |

TABLE 11: Explanatory independent variables for experimental design for tonotactics in Thlantlang Lai (Hyman, 2007, (18)).

| Tone | Noun | Word |
|---|---|---|
| HL | N1 | râal 'enemy' |
| | N2 | zôong 'monkey' |
| H | N1 | kóoy 'friend' |
| | N2 | vók 'pig' |
| L | N1 | bòoy 'chief' |
| | N2 | ràng 'horse' |

TABLE 12: Wordbank for $N_1 + N_2$ data set for Thlantlang Lai tonotactics (Hyman, 2007, (19)).

Because there are multiple explanatory variables in this design, another parameter of the design is how to investigate the interactions of the explanatory variables. The usual choice in fieldwork in this situation is to investigate all possible ways the explanatory variables may vary together—to choose a *factorial* design, where each level of each explanatory variable is cross-classified with another, e.g. since there are three tonal classes in Thlantlang Lai, there are 3 × 3 = 9 possible factor combinations (*treatments*) for a bigram, as shown in Table 13. This factorial design is exactly the same design that was used in cross-classifying all putative five tonal classes in Kirikiri prenominal possessive phrases (Table 6 in §2.3.3).

| | *+HL* | *+H* | *+L* |
|---|---|---|---|
| HL | HL *+HL* | HL *+H* | HL *+L* |
| H | H *+HL* | H *+H* | H *+L* |
| L | L *+HL* | L *+H* | L *+L* |

TABLE 13: Factorial design for underlying tone sequences over bigrams $N_1 + N_2$ in Thlantlang Lai (Hyman, 2007, (18)). The levels of the two explanatory variables, $N_1$ TONE and $N_2$ TONE, are cross-classified so that all possible combinations of tones (3×3=9 in total) are included. The levels of $N_2$ TONE are italicized so that they are distinguishable from those of $N_1$ TONE.

Hyman (2007)'s factorial experimental design for Thlantlang Lai tonotactics is not at all unusual in tonal fieldwork. For instance, Hyman (1985) examined the same kinds of constructions with factorial experimental designs in Bamileke-Dschang (Niger-Congo Bantoid, Cameroon); see Table 1 in (Hyman 1985), which lists disyllabic $N_1 + N_2$ associative constructions, e.g. *èfɔ̀ mə̀ndzwì* 'chief of leopards' (see also Bird 1999 for an extension of this work). The factorial design in testing all possible sequences of *n*-grams for some *n* is used all the time in production experiments on phonetic and phonological tonal sandhi, e.g. Xu (1997).

In fact, *any paradigmatic elicitation typically has a factorial design*. Consider the elicitation of verbal morphology: possible explanatory factors and their corresponding levels might be TENSE (past, present, future), PERSON (first, second, third), and NUMBER (singular, plural). To fill out the paradigm for a verb, we'd do a factorial experiment of TENSE × PERSON × NUMBER, for 3 × 3 × 2 = 18 possible treatments, one for each table cell in the factorial experimental design shown in Table 14.

| | PAST TENSE | |
|---|---|---|
| | **Number** | |
| | *Singular* | *Plural* |
| 1st | 1sg.past | 1pl.past |
| 2nd | 2sg.past | 2pl.past |
| 3rd | 3sg.past | 3pl.past |
| | PRESENT TENSE | |
| | **Number** | |
| | *Singular* | *Plural* |
| 1st | 1sg.pres | 1pl.pres |
| 2nd | 2sg.pres | 2pl.pres |
| 3rd | 3sg.pres | 3pl.pres |
| | FUTURE TENSE | |
| | **Number** | |
| | *Singular* | *Plural* |
| 1st | 1sg.fut | 1pl.fut |
| 2nd | 2sg.fut | 2pl.fut |
| 3rd | 3sg.fut | 3pl.fut |

TABLE 14: Factorial experimental design in verbal morphology elicitation for the set of explanatory variables TENSE (past, present, future), PERSON (first, second, third), and NUMBER (singular, plural). The design for the variable interaction is factorial since the variables are fully cross-classified and we have 3 (TENSE) × 3 (PERSON) × 2 (NUMBER) = 18 treatments.

**3 CONCLUSION.** This paper has shown how principles of experimental design can be used in designing structured elicitations in the iterative cycle of hypothesis generation and testing in fieldwork. The illustrative case studies presented in the paper focused on the early stages of toneme discovery. First, the paper showed that Pike (1948)'s classic toneme discovery procedure, described in §2.1, is a special case of the application of general principles of experimental design (Fisher 1925, 1935). Pike's discovery procedure was divided into two stages: (1) providing initial evidence for the existence of lexical tonal classes in a language, and (2) providing further evidence for lexical tonal contrasts by eliciting items from different putative tonal classes in controlled contexts. These two stages were illustrated with examples from Kirikiri elicitation data (§2.2.1, §2.3.3).

Section 2.2 described the first stage as an attempt to explain away variability in the pitch contour by acknowledging all known primary sources of influence on the pitch contour other than the hidden structure of tonal classes. If the attempt fails and there is a large amount of remaining unexplained variability, we then hypothesize that we are failing to take into account the latent variable of TONAL CLASS. Section 2.3 explicated the second stage of toneme discovery; it described manipulating the explanatory variable of TONAL CLASS while using strategies to control for the influence of confounding variables—variables that could potentially obscure the relation between TONAL CLASS and the pitch contour. The strategies include structuring the elicitation into blocks that are homogeneous with respect to confounding variables, e.g. with LENGTH of the word fixed at two syllables within a block, as well as randomizing the order of elicitation items within a block. Finally, Section 2.4 showed that the application of experimental design in structuring elicitations can be generalized beyond toneme discovery. Some practical aspects of elicitation such as making recordings, organizing elicitation items and associated files, and annotating phonetic data are also addressed in on-line tutorials available at http://www.krisyu.org/blog/posts/2013/06/ldc-kiy-overview/.

If the principles of experimental design that form the basis of an "experimental state of mind" illustrated in the paper seem very familiar to the fieldworker, it is because an experimental state of mind has always been a strong component of the fieldworker's state of mind. It's no accident that Pike (1948)'s toneme discovery procedure can be understood in terms of experimental design. It's also no accident that the essential qualities of the fieldworker's state of mind, of the love of discovery of language-particular uniqueness and dedication to "whole language" highlighted in Hyman (1985, p. 29-30), are also essential qualities of the experimental state of mind. This combination of meticulous attention to detail and embracing of the big picture is requisite for carefully addressing explanatory and confounding variables in designing structured elicitation.

## References

Andruski, Jean E. 2006. Tone clarity in mixed pitch/phonation-type tones. *Journal of Phonetics* 34. 388–404.

Baayen, R. H. 2008. *Analyzing linguistic data: a practical introduction to statistics*. Cambridge: Cambridge University Press.

Babel, Molly, and Dasha Bulatov. 2012. The role of fundamental frequency in phonetic accommodation. *Language and Speech* 55. 231–248.

Baken, R. J., and Robert F. Orlikoff. 2000. *Clinical measurement of speech and voice*. San Diego, CA: Singular Thomson Learning, 2nd edition.

Belotel-Grenié, Agnès, and Michel Grenié. 1997. Types de phonation et tons en chinois standard. Cahiers de Linguistique - Asie Orientale 26. 249–279.

Bird, Steven. 1999. Multidimensional exploration of online linguistic field data. *NELS 29*. 33-50.

Bird, Steven and Haejoong Lee. 2014. Computational support for early elicitation and classification of tone. *Language Documentation and Conservation* 8. 453–461

Boersma, Paul, and David Weenink. 2010. Praat: doing phonetics by computer (version 5.1.32) [computer program]. http://www.praat.org.

Brindley, G.S. 1960. *Physiology of the retina and the visual pathway*. London: Edward Arnold Ltd.

Chen, Matthew Y. 2000. *Tone sandhi*. Cambridge: Cambridge University Press.

Connell, Bruce. 2001. Downdrift, downstep, and declination. *Proceedings of Typology of African Prosodic Systems Workshop*.

Connell, Bruce. 2002. Tone languages and the universality of intrinsic $f_0$: evidence from Africa. *Journal of Phonetics* 30. 101–129.

Connell, Bruce, and D. Robert Ladd. 1990. Aspects of Pitch Realisation in Yoruba. *Phonology* 7. 1–29.

Deutsch, Diana, Trevor Henthorn, and Mark Dolson. 2004. Absolute pitch, speech, and tone language: some experiments and a proposed framework. *Music perception* 21. 339–356.

Eady, Stephen J., William E. Cooper, Gayle V. Klouda, Pamela R. Mueller, and Dan W. Lotts. 1986. Acoustical characteristics of sentential focus: narrow vs. broad and single vs. dual focus environments. *Language and Speech* 29. 233–280.

Fernald, Anne, and Thomas Simon. 1984. Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* 20. 104–113.

Fisher, Ronald A. 1925. *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.

Fisher, Sir Ronald A. 1935. *The design of experiments*. New York: Hafner Publishing Company, (9th edition 1971) edition.

Fougeron, Cécile, and Sun-Ah Jun. 1998. Rate effects on French intonation: prosodic organization and phonetic realization. *Journal of Phonetics* 26. 45–69.

Gandour, Jack, Apiluck Turntavitikul, and Nakarin Satthamnuwong. 1999. Effects of speaking rate on Thai tones. *Phonetica* 56. 123–134.

Gussenhoven, Carlos. 2004. *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Hayes, Bruce. 1989. The prosodic hierarchy in meter. In Paul Kiparsky and Gilbert Youmans (ed.), *Rhythm and meter*, 201–260. Orlando, FL: Academic Press.

Hombert, Jean-Marie. 1978. Consonant types, vowel quality, and tone. In Victoria A. Fromkin (ed.), *Tone: a linguistic survey*, 77–111. New York: Academic Press.

House, David. 1990. *Tonal perception in speech*. Lund, Sweden: Lund University Press.

Howie, John Marshall. 1976. *Acoustical studies of Mandarin vowels and tones*. Cambridge: Cambridge University Press.

Hyman, Larry M. 1985. Word domains and downstep in Bamileke-Dschang. *Phonology Yearbook* 2:47–83.

Hyman, Larry M. 2001. Fieldwork as a state of mind. In *Linguistic fieldwork*, ed. Paul Newman and Martha Ratliff, 15–33. Cambridge, UK: Cambridge University Press.

Hyman, Larry M. 2007. Elicitation as experimental phonology: Thlantlang Lai tonology. In Maria-Josep Solé, Patrice Speeter Beddor, and Manjari Ohala (ed.), *Experimental approaches to phonology*, 7–24. Oxford; New York: Oxford University Press.

Hyman, Larry M. 2011. Tone: Is it different? In John Goldsmith, Jason Riggle, and Alan C. L. Yu (ed.), *The handbook of phonological theory*, 197–239. Malden, MA and Oxford, Wiley-Blackwell.

Jaeger, T. Florian, Elisabeth Norcliffe, and Alice C. Harris (ed.). To appear. Special issue of Language, Cognition and Neuroscience.

Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA: Blackwell Publishing.

Katz, Jonah, and Elisabeth Selkirk. 2011. Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language* 87. 771–816.

Krishnan, Ananthanarayan, and Jackson T. Gandour. 2009. The role of the auditory brainstem in processing linguistically-relevant pitch patterns. *Brain and Language* 110. 135–148.

Krishnan, Ananthanarayan, Yisheng Xu, Jackson Gandour, and Peter Cariani. 2005. Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research* 25. 161–168.

Kuo, Yu-ching, Yi Xu, and Moira Yip. 2007. The phonetics and phonology of apparent cases of iterative change in Standard Chinese. In Carlos Gussenhoven and Tomas Riad (ed.), *Tones and tunes: Experimental studies in word and sentence prosody*, volume 2, 211–235. Berlin, Germany: Mouton de Gruyter.

Ladefoged, Peter. 2003. *Phonetic data analysis*. Malden, MA: Blackwell Publishing.

Liberman, Mark, and Janet Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In *Language sound structure*, 157–233. The MIT Press.

Maddieson, Ian. 1978. The frequency of tones. *UCLA Working Papers in Phonetics* 41. 43–52.

Mattock, Karen, and Denis Burnham. 2006. Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy* 10. 241-265.

Mattock, Karen, Monika Molnar, Linda Polka, and Denis Burnham. 2008. The developmental course of lexical tone perception in the first year of life. *Cognition* 106. 1367–1381.

Maxwell, Scott E., and Harold D. Delaney. 2004. *Designing experiments and analyzing data: a model comparison perspective*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2nd edition.

Montgomery, John C. 2005. *Design and analysis of experiments*. John Wiley & Sons, Inc., 6th edition.

Moore, Brian C. J. 2012. *An introduction to the psychology of hearing*. Bingley, UK: Emerald Group Publishing Limited, 6th edition.

Pike, Kenneth L. 1948. *Tone languages*. University of Michigan, Ann Arbor.

Plato. 360 B.C.E. *Phaedrus*. http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0174%3Atext%3DPhaedrus%3Asection%3D265e.

Remijsen, Bert. 2014. The study of tone in languages with a quantity contrast. *Language Documentation and Conservation* 8. 672–689.

Rietveld, Toni, and Roeland van Hout. 2005. *Statistics in language research: analysis of variance*. Berlin: Mouton de Gruyter.

Saville, D. J., and G. R. Wood. 1986. A method for teaching statistics using N-Dimensional geometry. *The American Statistician* 40. 205–214.

Shattuck-Hufnagel, Stefanie, and Alice E. Turk. 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* 25. 193–247.

Snider, Keith. 2014. On Establishing Underlying Tonal Contrast. *Language Documentation and Conservation* 8. 707–737.

Stanford, James N. 2010. The role of marriage in linguistic contact and variation: Two Hmong dialects in Texas. *Journal of Sociolinguistics* 14. 89–115.

Stern, Daniel N., Susan Spieker, and Kristine MacKain. 1982. Intonation contours as signals in maternal speech. *Developmental Psychology* 18. 727–735.

Sun, Xuejing, and Yi Xu. 2002. Perceived pitch of synthesized voice with alternate cycles. *Journal of Voice* 16. 443–459.

Talkin, David. 1995. A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn and K. K. Paliwal (ed.), *Speech coding and synthesis*, 495–518. Elsevier Science Inc.

Teller, Davida Y. 1984. Linking propositions. *Vision Research* 24. 1233–1246.

Whalen, D. H., and Yi Xu. 1992. Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica* 49. 25–47.

Wong, Patrick C. M., and Randy L. Diehl. 2003. Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language & Hearing Research* 46. 413–421.

Xu, Yi. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25. 61–83.

Xu, Yi. 1999. Effects of tone and focus on the formation and alignment of $f_0$ contours. *Journal of Phonetics* 27. 55–105.

Zhao, Yuan, and Dan Jurafsky. 2009. The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics* 37. 231–247.

Zsiga, Elizabeth, and Rattima Nitisaroj. 2007. Tone features, tone perception, and peak alignment in Thai. *Language and Speech* 50. 343–383.

Kristine M. Yu
krisyu@linguist.umass.edu

# APPENDIX

## A Elicitation item lists for Kirikiri

| Item | Target | Utterance | Gloss | Lexical class | Frame | Length (syll) |
|---|---|---|---|---|---|---|
| 1 | kɔɔ | kɔɔ | black | adj | Iso | 1 |
| 2 | kee | kee kɔɔ | black ant | n | adj-black | 1 |
| 3 | fɔɔ | fɔɔ kɔɔ | black wallaby | n | adj-black | 1 |
| 4 | kuu | kuu | female | adj | Iso | 1 |
| 5 | kee | kee kuu | female ant | n | adj-female | 1 |
| 6 | fɔɔ | fɔɔ kuu | female wallaby | n | adj-female | 1 |
| 7 | sɔɔ | sɔɔ | small | adj | Iso | 1 |
| 8 | kee | kee sɔɔ | small ant | n | adj-small | 1 |
| 9 | fɔɔ | fɔɔ sɔɔ | small wallaby | n | adj-small | 1 |
| 10 | kee | kee | ant | n | Iso | 1 |
| 11 | fɔɔ | fɔɔ | wallaby | n | Iso | 1 |
| 12 | fiɣɜɜ | fiɣɜɜ | large | adj | Iso | 2 |
| 13 | kee | kee fiɣɜɜ | large ant | n | adj-large | 1 |
| 14 | fɔɔ | fɔɔ fiɣɜɜ | large wallaby | n | adj-large | 1 |
| 15 | sije | sije taɾʉ | the pig sleeps | n | vp-sleep | 2 |
| 16 | naβi | naβi taɾʉ | the dog is sleeping | n | vp-sleep | 2 |
| 17 | ɸaɰai | ɸaɰai taɾʉ | the bandicoot is sleeping | n | vp-sleep | 2 |
| 18 | fɔɔ | fɔɔ taɾʉ | the wallaby is sleeping | n | vp-sleep | 1 |
| 19 | kaza | kaza taɾʉ | the gecko is sleeping | n | vp-sleep | 2 |
| 20 | sije | sije | pig | n | iso | 2 |
| 21 | naβi | naβi | dog | n | iso | 2 |
| 22 | ɸaɰai | ɸaɰai | bandicoot | n | iso | 2 |
| 23 | fɔɔ | fɔɔ | wallaby | n | iso | 1 |
| 24 | kaza | kaza | gecko | n | iso | 2 |
| 25 | sije | sije kʍaa zaɾe | the pig is making a sound | n | vp-sound | 2 |
| 26 | naβi | naβi kʍaa zaɾe | the dog is making a sound | n | vp-sound | 2 |
| 27 | ɸaɰai | ɸaɰai kʍaa zaɾe | the bandicoot is making a sound | n | vp-sound | 2 |
| 28 | fɔɔ | fɔɔ kʍaa zaɾe | the wallaby is making a sound | n | vp-sound | 1 |
| 29 | kaza | kaza kʍaa zaɾe | the gecko is making a sound | n | vp-sound | 2 |
| 30 | krʉʉ | krʉʉ kʍaa zaɾe | the frog is making a sound | n | vp-sound | 1 |
| 31 | kee | kee kʍaa zaɾe | the ant is making a sound | n | vp-sound | 1 |
| 32 | fuu | fuu kʍaa zaɾe | the honey bee is making a sound | n | vp-sound | 1 |
| 33 | tiɾʊ | tiɾʊ kʍaa zaɾe | the kingfisher is making a sound | n | vp-sound | 2 |
| 34 | krʉʉ | krʉʉ | frog | n | iso | 1 |
| 35 | fuu | fuu | honey bee | n | iso | 1 |
| 36 | tiɾʊ | tiɾʊ | kingfisher | n | iso | 2 |

TABLE A.1: List of elicitation items for kiy-20111208-6-ap-nps-vps.

| Item | Kirikiri | Gloss | $N_1$ tone | $N_2$ tone | $N_1 + N_2$ tones | $N_1$ | $N_2$ |
|---|---|---|---|---|---|---|---|
| 1 | ɸaɰai giɰʉ | bandicoot's elbow | 1 | 1 | 11 | ɸaɰai | giɰʉ |
| 2 | kaza giɰʉ | gecko's elbow | 2 | 1 | 21 | kaza | giɰʉ |
| 3 | ɸiʋa giɰʉ | snail's elbow | 3 | 1 | 31 | ɸiʋa | giɰʉ |
| 4 | nala giɰʉ | wasp's elbow | 4 | 1 | 41 | nara | giɰʉ |
| 5 | taʋa giɰʉ | catfish's elbow | 5 | 1 | 51 | taʋa | giɰʉ |
| 6 | ɸaɰai oɾa | bandicoot's tongue | 1 | 2 | 12 | ɸaɰai | oɾa |
| 7 | kaza oɾa | gecko's tongue | 2 | 2 | 22 | kaza | oɾa |
| 8 | ɸiʋa oɾa | snail's tongue | 3 | 2 | 32 | ɸiʋa | oɾa |
| 9 | nala oɾa | wasp's tongue | 4 | 2 | 42 | nara | oɾa |
| 10 | taʋa oɾa | catfish's tongue | 5 | 2 | 52 | taʋa | oɾa |
| 11 | ɸaɰai ɸaɰao | bandicoot's groin | 1 | 3 | 13 | ɸaɰai | ɸaɰao |
| 12 | kaza ɸaɰao | gecko's groin | 2 | 3 | 23 | kaza | ɸaɰao |
| 13 | ɸiʋa ɸaɰao | snail's groin | 3 | 3 | 33 | ɸiʋa | ɸaɰao |
| 14 | nala ɸaɰao | wasp's groin | 4 | 3 | 43 | nara | ɸaɰao |
| 15 | taʋa ɸaɰao | catfish's groin | 5 | 3 | 53 | taʋa | ɸaɰao |
| 16 | ɸaɰai kʍawaa | bandicoot's chin | 1 | 4 | 14 | ɸaɰai | kʍawaa |
| 17 | kaza kʍawaa | gecko's chin | 2 | 4 | 24 | kaza | kʍawaa |
| 18 | ɸiʋa kʍawaa | snail's chin | 3 | 4 | 34 | ɸiʋa | kʍawaa |
| 19 | nala kʍawaa | wasp's chin | 4 | 4 | 44 | nara | kʍawaa |
| 20 | taʋa kʍawaa | catfish's chin | 5 | 4 | 54 | taʋa | kʍawaa |
| 21 | ɸaɰai kʊɾee | bandicoot's string | 1 | 5 | 15 | ɸaɰai | kʊɾee |
| 22 | kaza kʊɾee | gecko's string | 2 | 5 | 25 | kaza | kʊɾee |
| 23 | ɸiʋa kʊɾee | snail's string | 3 | 5 | 35 | ɸiʋa | kʊɾee |
| 24 | nala kʊɾee | wasp's string | 4 | 5 | 45 | nara | kʊɾee |
| 25 | taʋa kʊɾee | catfish's string | 5 | 5 | 55 | taʋa | kʊɾee |

TABLE A.2: List of elicitation items for kiy-20111213-1-kiy-ap-framedwordlist.

## B Sources of influence on fundamental frequency and pitch

| Variable | Example levels | Sample references |
|---|---|---|
| Speech rate | slow, fast | Fougeron and Jun (1998); Gandour et al. (1999); Kuo et al. (2007) |
| Sex | male, female | Baken and Orlikoff (2000, p. 175-176) |
| Age | 21-26, 85 years | Baken and Orlikoff (2000, p. 173-176) |
| Speaker pitch range | small, large | Baken and Orlikoff (2000, p. 175-176) |
| Background noise | absent, present | Zhao and Jurafsky (2009) |
| Consonant voicing | voiced, unvoiced | Hombert (1978) |
| Vowel quality | high, low | Connell (2002); Hombert (1978) |
| Vowel length | short, long | Remijsen (2014) |
| Stress | stressed, unstressed | Remijsen (2014 |
| Prosodic position | utterance-final, phrase-medial, isolation | Maddieson (1978, p. 45-46), Hayes (1989), Shattuck-Hufnagel and Turk (1996), Zsiga and Nitisaroj (2007) |
| Declination | phrase-initial, phrase-final | Connell and Ladd (1990), Connell (2001), Gussenhoven (2004, Ch. 6) |
| Downstep | post-trigger | Connell and Ladd (1990), Connell (2001), Gussenhoven (2004, Ch. 6) |
| Preceding f0 for relative pitch perception | low, high | Wong and Diehl (2003), Huang and Holt (2009) |
| Tonal coarticulation and sandhi | preceding H, following L | Xu (1997); Chen (2000); Kuo et al. (2007) |
| Lexical frequency | low, high | Zhao and Jurafsky (2009) |
| Person | 1sg, 3sg | Hyman (2011, p. 203) |
| Tense/aspect | present, past | Hyman (2011, p. 203) |
| Negation | present, absent | Hyman (2011, p. 203) |
| Case | nominative, accusative | Hyman (2011, p. 203) |
| Emphasis level | 1 (mumble), 10 (shout) | Liberman and Pierrehumbert (1984) |
| Focus | contrastive focus, out-of-the-blue focus | Eady et al. (1986), Xu (1999), Katz and Selkirk (2011) |
| Giveness | given, new | Katz and Selkirk (2011) |
| Speech style | spontaneous, reading | Fernald and Simon (1984), Baken and Orlikoff (2000, p. 175-176) |
| Phonetic accommodation | pre-exposure, post-exposure | Babel and Bulatov (2012) |
| Modulation of attention | eliciting attention, maintaining attention in infant-directed speech | Stern et al. (1982) |
| Community identity | loyalty to husband's community, loyalty to father's community | Stanford (2010) |

Table B.1: Some additional examples of potential confounding variables in toneme discovery, arranged roughly in order from variables related to: physiology and the speech signal, phonetics and phonology, morphosyntax, and discourse and pragmatics, and social factors.