A Psycholinguistic Tool for the Assessment of Language Loss: The HALA Project

William O'Grady
Amy J. Schafer
Jawee Perla
On-Soon Lee
Julia Wieting

University of Hawai'i at Mānoa

A major obstacle to the early diagnosis of language loss and to the assessment of language maintenance efforts is the absence of an easy-to-use psycholinguistic measure of language strength. In this paper, we describe and discuss a body-part naming task being developed as part of the Hawai'i Assessment of Language Access (HALA) project. This task, like the others in the HALA inventory, exploits the fact that the speed with which bilingual speakers access lexical items and structure-building operations in their two languages offers a sensitive measure of relative language strength. In a pilot study conducted with Korean-English bilinguals, we were able to establish a strong correlation between language strength and naming times even in highly fluent bilingual speakers, in support of the central assumption underlying the HALA tests. We discuss the implications of this finding for the broader study of language strength as well as for the practical problems associated with work on language loss, maintenance, and revitalization.

1. INTRODUCTION[1]. It seems safe to assume that there is no such thing as a natural inclination to abandon one's native language. When a community shifts to a new language, it is always in response to external economic, social and political pressures (e.g., Nettle and Romaine 2000). This notwithstanding, language loss is ultimately a neurological phenomenon. Of necessity, it involves changes to the words, structure-building operations, and other resources that are implemented in the brain as "language" and employed in the course of communication. As we will show in this paper, this simple fact opens the door to the psycholinguistic assessment of language loss in individuals and in communities, offering researchers new tools for tracking this phenomenon and even for measuring the effects of language revitalization and maintenance programs.

We begin with a brief discussion of what it means to be proficient in a language and how the demands of proficiency increase with bilingualism—the usual precursor to language weakening and loss. We then introduce a project that we have undertaken to assess the relative strength of particular pairs of languages in bilinguals, and report on the results that we have obtained in a preliminary series of experiments. We conclude with some remarks about the possible usefulness of this type of work for the study of language loss and language revitalization.

2. BILINGUALISM AND LANGUAGE MAINTENANCE. Proficiency in a language involves access to a lexicon containing tens of thousands of words and to a set of routines for combining those words into phrases and sentences. Maintenance of such an intricate system presents very significant challenges. De Bot (2004:234) puts it this way:

> ... all the languages in the system need maintenance and advanced use ... It's not about how much memory space we have to store language material, since there probably is no real limit there, but about the time and resources needed to keep all parts of the system in the foreground of processing … learning another language does not remove older languages from memory, but does push them more to the background and makes it accordingly more difficult to access them.

The maintenance of two language systems at comparable levels of activation—the sort of bilingual state that staves off language loss—is no easy task. As Jessner (2003:241) notes, "psycholinguistic systems containing two or more language systems" are "less stable than monolingual ones, and repair or reactivation procedures are constantly required to maintain the system in a steady state."

The factor that contributes most directly to the maintenance of a linguistic system is the frequency with which it is used. Put simply, the more often the words and structure-building routines of a particular language are activated, the more accessible they are. And of course, the more accessible the system is, the more likely speakers are to feel comfortable using it. There is a natural cycle here: as a language becomes less accessible through infrequent use, its speakers become reluctant to use it, further decreasing its accessibility and creating the downward spiral that ultimately leads to language loss.
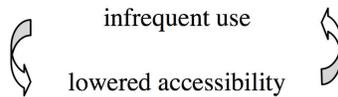
infrequent use

lowered accessibility

FIGURE 1: The cycle of decreasing usage and lowered accessibility that leads to language loss

A widely acknowledged psycholinguistic reflex of accessibility is speed—a more highly activated lexical item or structure-building routine is accessed more quickly than a less highly activated counterpart. Thus, as illustrated below, frequency of use translates into higher activation or strength, which makes possible quicker access.

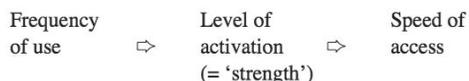| Frequency of use | ⇨ | Level of activation (= 'strength') | ⇨ | Speed of access |

FIGURE 2. Usage, activation, and speed of access.

The speed with which a speaker can access the vocabulary items and structure-building routines of a language thus serves as a potent indicator of the system's level of activation.

The theoretical claims underlying this scenario make up what is sometimes referred to as the "Weaker Links Hypothesis" (e.g., Gollan et al. 2008): the infrequent use of a language leads to a weakening of the associations between forms and their meanings, which in turn is reflected in lower levels of activation and slower access times. As we will see next, this idea opens the door for the development of a simple psycholinguistic measure of language strength and language shift—the principal objective of the Hawai'i Assessment of Language Access (HALA[2]) project, to which we now turn.

3. THE HALA PROJECT—AN EXPERIMENT. The measure on which the HALA project focuses is a comparative one—speed of access to words and structure-building operations in one language *relative to a speaker's other language(s)*. Thus, it does not matter whether speaker A is faster at accessing the word for 'nose' in, say, Chamorro than is speaker B. What matters is whether speaker A is faster at accessing the word for 'nose' in English than in Chamorro, or vice versa. It is asymmetries of this type that can ultimately serve as indicators of language strength. We will illustrate this point with the help of a lexical access test involving body part terms—one of the inventory of tasks in the HALA project.

Our idea in devising the body-part naming test was to focus on a semantic field with the following three properties.

- It includes words for which we can expect counterparts in all languages, as evidenced by the importance of basic body part terms to work in comparative and historical linguistics. At least some of the words in question are basic enough to have been acquired by all users of the language at an early age. Thus, evidence of poor or slow access should be a highly reliable indicator of language weakening.

---

[2] By happy coincidence, *hala* is the Hawaiian name for 'pandanus', a tree found on many Pacific islands. Its leaves are commonly used for weaving in Hawai'i and elsewhere.

- Because of their basic status, body-part terms can also be expected to be relatively resistant to replacement by borrowing. As such, we can reasonably expect elicitation of those terms to result in the production of words from the target language rather than items borrowed from a competitor language, as might happen if, for instance, we elicited items referring to electronic devices.

A pilot study involving eleven highly bilingual speakers of English and Korean helps illustrate the effectiveness of the body-part naming test and the logic underlying the HALA project.

3.1 PARTICIPANTS. All of our participants had been born in the United States and had been exposed to both English and Korean from birth. All considered English to be their stronger language and reported that Korean constituted between 10 and 50% of their daily language use (mean = 35%). The participants were all graduate or undergraduate students at the University of Hawai'i at Mānoa, and ranged in age from 19 to 27 years old.

We had a two-fold motivation for conducting our pilot study with Korean-English bilinguals. First, these participants, who were readily available to us, were similar to speakers of endangered languages in a crucial respect—they had been exposed to a family language (Korean) at home and to a more widely spoken competitor language (English) outside the home. Second, we had access to independent assessments of the proficiency of these speakers in their two languages—an essential prerequisite for evaluating the accuracy of our test.

3.2 MATERIALS. The implementation of the body-part naming test is extremely simple. Speakers name body parts in response to a series of photographs (see samples in figures 3 and 4), naming times are measured in milliseconds from the onset of the photo to the onset of the response, and these times are compared for the two languages of interest.

FIGURE 3: "Eye"   FIGURE 4: "Eyebrow"

There were a total of 43 test items, divided into three subsets or strata based on their relative frequency of use, as determined by information collected from intuitive ratings, naming times and HAL log frequency[3] in the English Lexicon Project (Balota et al. 2007), and performance by a separate group of pilot participants who spoke a range of native languages. The items are listed in table 1.

High frequency	Medium frequency	Low frequency
back	arm	ankle
ear	cheek	arch
eye	chin	bicep
face	eyebrow	calf
fingers	fingernail	cheekbone
foot	forehead	elbow
hair	neck	eyelid
hand	palm	forearm
head	thumb	heel
knee	toe	knuckle
leg	waist	pupil
lips	wrist	shin
mouth		toenails
nose		
shoulder		
stomach		
teeth		
tongue		

Table 1: Test items by stratum.

Differences in frequency across languages cannot be entirely avoided, of course. As an anonymous reviewer notes, for instance, the frequency of a word such as 'forearm' might well be higher than otherwise expected if the lexical item in question is also used for 'arm' or even 'hand/arm,' as happens in some languages. However, the effect of this variation can be minimized by choosing (as we did) referents that are likely to be of comparable relative relevance in all communities (e.g., there are presumably more references to faces than to

---

[3] HAL log frequency values are log-transformed frequencies from the HAL corpus, which consists of approximately 131 million words collected from Usenet newsgroups (Lund and Burgess 1996).

elbows in all languages). Interestingly, Bates et al. (2003) report strong cross-language cor-relations in frequency and naming times for the seven languages that they examined, also noting that factors such as word length, syllable structure, and morphological composition are less stable and less important than frequency and conceptual familiarity in predicting naming times. We can therefore expect the HALA task to provide at least a good first ap-proximation of differences in relative language strength.[4]

3.3 DESIGN AND PROCEDURE. Each participant was tested in both languages. One can expect that naming times will be shorter on the second run through the test, so we balanced the testing order between participants. Half were tested first in Korean, and then in English, while the other half received the reverse order.

Each testing session began with simple instructions and a set of 12 practice items so that the speakers could become accustomed to the task. The main set of items were ordered so that the high-frequency subset always appeared first, followed by the medium-frequency subset and then the low-frequency subset. However, within each subset we provided a dif-ferent random order of the items for each language. The randomization within each subset minimized the likelihood that the participants would generate expectations in their second testing session about which item would appear next. In addition, some earlier piloting re-sults suggested that separating the items by strata facilitated the participants' progression from more basic vocabulary items to more specialized ones, making it easier for them to respond rapidly to each item in turn. One likely effect of this ordering was to make clear to participants that we were expecting the most basic term that applied to the depicted body part (e.g., *"arm,"* not *"appendage"* or *"limb"*).

For each item, a trial began with the onset of a photo, displayed in the center of a com-puter monitor in a quiet room. Each photo was a black and white image of an area of the body, in which the critical body part was encircled in red, as shown in figures 3 and 4. The onset of the photo was synchronized with a short beep, to draw the speaker's attention. In this version of the HALA test (we have also developed a more portable implementation), the photo remained on the screen until the participant responded by naming the item aloud or asking to skip the item. Naming times were recorded by a millisecond-accurate response box equipped with a voice key. Following the onset of the naming response, a version of the photo without the red circle remained on the screen for another 2000 ms, allowing the speakers time to complete their response and prepare to attend to the next item. The entire session was audio recorded so that inaccurate responses and other errors could later be eliminated.

---

[4] Our division into three strata provides some protection against the possibility of confounds. We make clear predictions that the effect of dominance should hold across all three strata, although per-haps to different degrees, as discussed further below. If the effect does not hold, the data then indicate that the items require more detailed analysis (such as adding the effect of differences in word length) in order to achieve appropriate matching between the pair of languages.

3.4 PREDICTIONS. Consistent with well-established psycholinguistic principles (Gollan et al. 2008 and the many references cited there), naming times are inversely correlated with frequency of use: high-frequency words have shorter naming times than low-frequency words. The stronger language thus produces, on average, shorter naming times than the weaker language. In addition, this effect increases as item frequency decreases, leading to the following predictions.

- A main effect of frequency, which also holds within each language: faster response times for more frequent words.
- A main effect of language strength: faster response times for the stronger language.
- An interaction between frequency and language strength: the language strength effect is greater for lower frequency words than for higher frequency words.

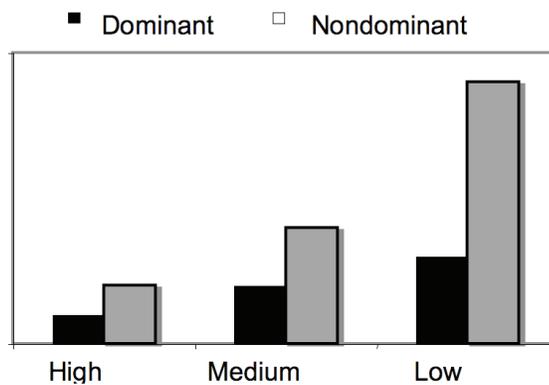The expected pattern of naming times is depicted in figure 5.

FIGURE 5: Expected pattern of naming times.

As can be seen here, we expect naming times to be shorter for high-frequency vocabulary items than for lower-frequency items, and we expect that items from the same stratum to have shorter naming times in the stronger language. These expectations were borne out.

3.5 RESULTS. Figure 6 summarizes the accuracy of our participants in responding to our picture stimuli—that is, the rate at which they correctly named each picture.

As can be seen here, the participants exhibit a very high level of accuracy in both languages on all three vocabulary strata, with no significant effect of language emerging overall or in any of the subsets of words in our test, but a slight numerical advantage for English. This confirms that our participants were in fact highly bilingual. The results are consistent with their self-assessment that English was their stronger language and the one
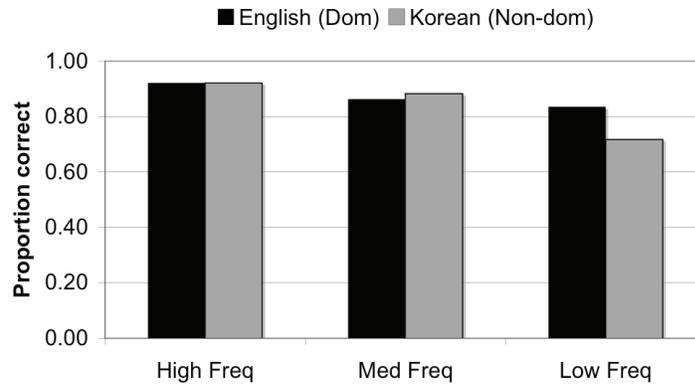
■ English (Dom) ■ Korean (Non-dom)

FIGURE 6: Accuracy on the naming task.

they have greater exposure to and use of. Statistically, a repeated measures analysis of variance (ANOVA) treating participants as a random variable found a significant effect of strata ($F(2,20) = 4.150$, $p < .05$), but not of language ($F < 1$) or of the interaction of strata and language ($F < 1.7$). This confirms that accuracy was higher for the more frequent items than the less frequent ones, but did not differ in any meaningful fashion between the two languages.

The calculation of naming times, the key measure in our task, was conducted for just those test items in which the stimulus picture had been correctly identified. As is common in psycholinguistic research, we also performed a simple screening to remove extreme values, eliminating any naming times for each participant that were more than 2.5 standard deviations from the overall mean naming time for accurate responses from that participant. The results of this calculation are presented in figure 7.

The key finding here is that our participants had significantly faster naming times for all three strata of vocabulary items in English, compared to Korean. In addition to confirming that English is the stronger language for our participants, this finding underlines one of the principal advantages of the HALA approach to the assessment of language strength: English emerges as the stronger language for all three subsets of vocabulary even though the participants are all highly fluent in Korean and even though no difference was evident on accuracy measures (see figure 6). Statistically, a repeated measures ANOVA[5] found significant effects of strata ($F(2,18) = 39.129$, $p < .01$), language ($F(1,9) = 36.879$, $p < .01$, and their interaction ($F(2,18) = 5.092$, $p < .05$). The effect of language was further verified by significant effects of language in paired t-tests within each stratum (all p's $< .02$). In other words, there were statistically reliable differences in naming times across strata

---

[5] One speaker was missing many values from the least frequent strata in Korean after elimination of inaccurate responses and outlying naming times, resulting in the loss of one degree of freedom in this analysis. However, the data patterns remain the same with other treatments, showing robust effects of language and strata.
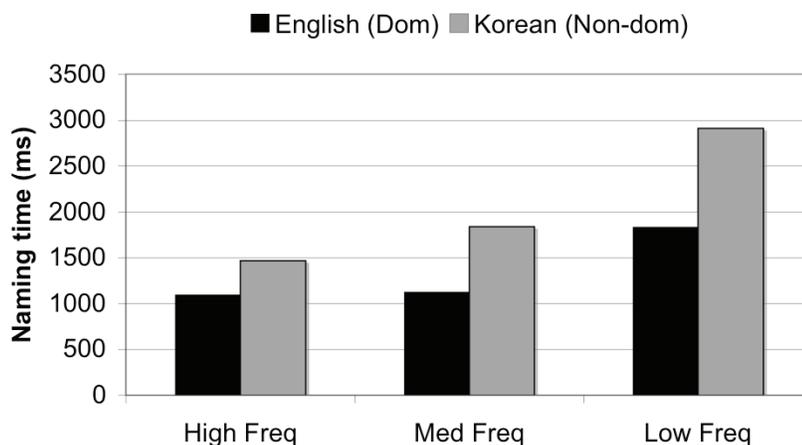
Figure 7: Naming times for accurate responses.

and between the two languages, and (as predicted) the effect of language strength varied across strata: the effect was significant for each subset of items, and strongest for the least frequent items.

As a further probe of the validity of our test, we calculated the naming times for the five subjects who had the highest rates of naming accuracy—at least 90% correct across the three subsets of vocabulary. These participants, like the larger set of participants, showed no effect of language on accuracy (F(1,4) = 1.719, p = .26), and actually showed slightly higher accuracy in Korean (96% correct) than in English (92% correct), although their self-assessments as well as independent assessments agree that English is their stronger language. The results are presented in Figure 8.

As can be seen here, the higher accessibility of English is still strongly evident, with significantly shorter naming times for that language. Statistically, we once again found significant effects of strata (F(2,8) = 36.676, p < .01, language (F(1,4) = 18.673, p < .05, and their interaction (F(2,8) = 6.300, p < .05). Paired t-tests within each strata showed marginal effects for the high- and medium-frequency sets (t's = 2.4, p's = .07) and a significant effect for the lowest frequency word set (t = 8.16, p < .01).

3.6 DISCUSSION. The results from our body-part test support three findings. First, accuracy declined with decreasing frequency, but did not show reliable effects of language strength. This demonstrates that although accuracy can be a useful measure of language strength, it is less sensitive than desired for highly bilingual populations that might have subtle differences in the relative strength of their two languages.

Second, consistent with independently established psycholinguistic principles, naming times (our key measure) show significant effects for both frequency and language strength, and for their interaction. Thus our participants responded faster to more frequent stimuli in both languages, but were overall faster on all subsets of vocabulary in their stronger language (English).
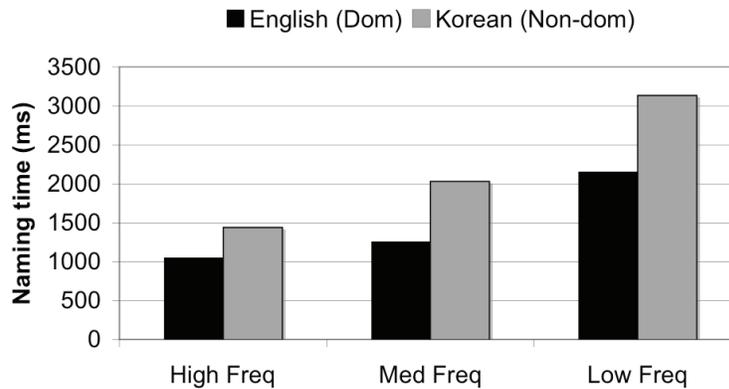
FIGURE 8: Naming times for the 5 subjects with the highest accuracy rates

Finally, the language strength effect remains significant even with highly accurate speakers. This confirms that naming times provide a sensitive and effective measure of strength, thereby buttressing the key assumptions underlying the HALA approach to language assessment.

Needless to say, we do not take these results to indicate that Korean is endangered, or even that the particular subjects who we tested will lose their ability to speak and understand Korean. Our goal has simply been to establish that a psycholinguistic test of language activation can provide extremely subtle measures of language strength, even in the case of speakers who seem to be highly bilingual. The interpretation of the sociolinguistic import of these measures will of course depend on a wide range of factors specific to particular groups of speakers and their languages.

4. CONCLUDING REMARKS. At first glance, the most obvious way to measure a language's strength would be to probe knowledge of specialized vocabulary (fish or plant names, for instance), intricate inflectional paradigms, complex structural patterns, register-related contrasts, and the like. However, such an approach encounters many obstacles. Not only do the test materials have to be tailored to each specific language, their formulation would require detailed knowledge of the language's workings. This is fundamentally impractical in the case of many languages, including almost all endangered languages, which are typically little studied in the first place.

Our idea is very different. The starting point is the simple observation that the mastery and maintenance of virtually all aspects of language, from vocabulary to morphosyntax, are sensitive to frequency of use, which in turn correlates with accessibility (strength). This in turn makes it possible to exploit another simple fact: accessibility is indexed by speed of access. We can thus get a good initial indication of a language's strength by measuring the speed with which speakers access its vocabulary and structure-building operations relative

to those of their other language(s). The test of body part terms outlined above illustrates this approach, showing how a very simple naming task can yield easy-to-interpret results about lexical access.

In addition to their empirical success, the tests that comprise the HALA approach to language assessment have a number of practical advantages.

- They are sensitive to even small differences in language strength.
- They are inexpensive.
- They can be used for any language.
- They are not dependent on orthography.
- The researcher need not have extensive knowledge of the languages being tested to begin the process of assessment.
- The testing time for each task is short.
- They can be used with children or adults.
- They are fully portable—they can be run on a personal computer with readily available software, together with an audio recorder. (In addition to the "laboratory" version implemented for the Korean-English bilinguals, we have also created a more portable implementation using video-based presentation.)
- The data can be collected by field workers, educators, and other researchers with no special training in psycholinguistics (or even linguistics). Basic analyses of the type reported above can be conducted with minimal training and readily available software by following the steps in our field manual.

In these respects, the HALA approach contrasts sharply with more traditional approaches to language assessment, including those that rely on language-specific tests of vocabulary such as the Bilingual Verbal Ability Test (Muñoz-Sandoval et al. 2005), measures of mean utterance length (Yip and Matthews 2006), and estimates of proportions of language use in particular speech situations (the Graded Intergenerational Interruption Scale of Fishman 1991, and the Language Maintenance Index of Lewis 1996). Compared to the HALA tests, such assessment instruments are highly labor-intensive, require extensive knowledge of the target languages, and often yield results whose interpretation is far from straightforward.

The next step in the development of the HALA project involves a series of extensive field tests, some of which have already been arranged. One of our goals is to conduct work in communities where independent information about language strength is already available, which will enable us to further "test our test." If we are on the right track, our results should confirm that one language is in fact weaker in relevant groups of speakers. Once our test is validated in this way, it will be possible to use it with increasing confidence to support various types of research, including:

- Identifying early signs of potential language endangerment (e.g., 80% of all teenagers manifest quicker access to the competitor language than to the indigenous language).

- Comparing language access in different communities (e.g., those in rural areas versus those in urban areas, those with access to international satellite TV versus those without such access, and so on).
- Comparing language access in different cohorts (fifth grade students in 2009 versus fifth grade students in 2015).
- Assessing the effectiveness of language conservation programs (students in schools with instruction in the indigenous language through eight grades compared to those with instruction in the indigenous language through six grades).
- Assessing the relative strength of a particular language in candidate consultants in the case of fieldwork situations.

As these potential applications illustrate, the HALA tests comprise a new mode of language documentation that can be used to supplement and support traditional techniques by creating a psycholinguistic record of a language's strength in the individuals and communities who speak it. This in turn opens the door for a new set of research initiatives, creating opportunities that can be expected to contribute in substantial ways to the field of language documentation and conservation.

Authors' note:
For information on the availability of the HALA test materials, go to:
http://www.ling.hawaii.edu/research/hala.html

REFERENCES

BALOTA, DAVID A., MELVIN J. YAP, MICHAEL J. CORTESE, KEITH A. HUTCHISON, BRETT KESSLER, BJORN LOFTIS, JAMES H. NEELY, DOUGLAS L. NELSON, GREG B. SIMPSON, AND REBECCA TREIMAN. 2007. The English Lexicon Project. *Behavior Research Methods* 39:445–459.

BATES, ELIZABETH, SIMONA D'AMICO, THOMAS JACOBSEN, ANNA SZÉKELY, ELANA ANDONOVA, ANTONELLA DEVESCOVI, DAN HERRON, CHING CHING LU, THOMAS PECHMANN, CSABA PLÉH, NICOLE WICHA, KARA FEDERMEIER, IRINI GERDJIKOVA, GABRIEL GUTIERREZ, DAISY HUNG, JEANNE HSU, GOWRI IYER, KATHERINE KOHNERT, TEODORA MEHOTCHEVA, ARACELI OROZCO-FIGUEROA, ANGELA TZENG, AND OVID TZENG. 2003. Timed picture naming in seven languages. *Psychonomic Bulletin & Review* 10:344-380.

DE BOT, KEES. 2004. Introduction: Special issue on language attrition. *International Journal of Bilingualism* 8:233–237.

FISHMAN, JOSHUA. 1991. *Reversing language shift*. Philadelphia: Multilingual Matters.

GOLLAN, TAMAR, ROSA MONTOYA, CYNTHIA CERA, and TIFFANY SANDOVAL. 2008. More use almost always means a smaller frequency effect: Aging, bilingualism, and the Weaker Links Hypothesis. *Journal of Memory and Language* 58:787–814.

JESSNER, ULRIKE. 2003. A dynamic approach to language attrition in multilingual systems. In *Effects of the second language on the first. (Second language acquisition 3)*, ed. by Vivian James Cook, 234–246. Clevedon, UK: Multilingual Matters.

LEWIS, M. PAUL. 1996. Language maintenance in seven K'iche' communities. *SIL Electronic Working Papers in Linguistics* 2. Available at http://www.sil.org/silewp/1996/002/silewp1996-002.html.

LUND, KEVIN, and CURT BURGESS. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28:203–208.

MUÑOZ-SANDOVAL, ANA, JIM CUMMINS, CRISELDA ALVARADO, and MARY RUEF. 2005. *Bilingual Verbal Ability Tests*. Boston: Riverside Publishing.

NETTLE, DANIEL, and SUSANNE ROMAINE. 2000. *Vanishing voices*. Oxford; New York: Oxford University Press.

YIP, VIRGINIA, and STEPHEN MATTHEWS. 2006. Assessing language dominance in bilingual acquisition: A case for mean length of utterance differentials. *Language Assessment Quarterly* 3:97–116.

William O'Grady
ogrady@hawaii.edu

Jawee Perla
perla@hawaii.edu

Amy J. Schafer
aschafer@hawaii.edu

On-Soon Lee
onsoon@hawaii.edu

Julia Wieting
wieting@hawaii.edu