Using the Web as a Phonological Corpus: a case study from Tagalog

Kie Zuraw
Department of Linguistics
UCLA
Los Angeles, U.S.A.
kie@ucla.edu

Abstract

Some languages' orthographic properties allow written data to be used for phonological research. This paper reports on an on-going project that uses a web-derived text corpus to study the phonology of Tagalog, a language for which large corpora are not otherwise available. Novel findings concerning the phenomenon of intervocalic tapping are discussed in detail, and an overview of other phonological phenomena in the language that can be investigated through written data is given.

1 Introduction

Because the field of phonology studies *sound* patterns of languages, corpus-based phonology typically relies on audio corpora. These are expensive to create, and usually must undergo laborious hand-tagging to be useful. For much phonological investigation, there is no way around these harsh facts. Sometimes, however, a language's phonology and orthography conspire to allow phonological data to be gleaned from text. Abigail Cohn and Lisa Lavoie (p.c.), for example, have used text data on English comparatives to determine whether words are treated as monosyllabic, taking suffixal *X-er*, or longer, taking periphrastic *more X*. The cases of interest are words such as *feel* and *fire*, which have a tense or diphthongal nucleus followed by *l* or *r*, and are felt by many English speakers to be longer than monosyllabic. Corpus data on the frequencies of the two comparative types can be used as further evidence on the status of such words.

The Tagalog language (Austronesian, Philippines) exhibits several morphophonological phenomena that are reflected in its spelling. All of these phenomena involve some variation, which makes them ideal for text-corpus study: only with large amounts of data can we investigate the distribution of the variants and search for the factors that condition the variation. See Schachter & Otanes (1972) for basic descriptions of most of these phenomena:

- intervocalic tapping (*d* can become the tap sound [ɾ], spelled *r*, when it is between two vowels): ***d**umi* 'dirt' *ma-**r**umi* 'dirty'
- vowel-height alternations (*o* in final syllables can alternate with *u* in non-final syllables; there is a similar but more complicated *i*/*e* alternation): *hal**o*** 'mix' *hal**u**-in* 'to be mixed'
- nasal assimilation (a nasal consonant can take on the place of articulation of a following consonant): *pa**m**-butas* 'borer' *pa**n**-damot* 'picker-upper' *pa**ng**-gamas* 'trowel' (*ng* represents the velar nasal [ŋ])
- nasal substitution (stem-initial obstruents can turn into nasals when certain prefixes are added): ***p**ili* 'choosing' *ma-**m**ili* 'to choose'
- syncope (the vowel of a stem's final syllable can be deleted when a suffix is added, and the consonants that consequently become adjacent can undergo changes): *gaw**a*** 'act' *gaw_-in* 'to be done', *ti**ngi**n* 'look' *ti**g**_n-an* 'to be looked at'
- partial reduplication (when foreign stems that begin with consonant sequences and/or foreign consonants such as *f* un-

dergo copying of the first syllable, the consonant sequence can be simplified and the foreign consonant can be nativized): *nag-**fri**-friendster* ~ *nag-**pi**-friendster* 'using Friendster'

- infix location (in foreign stems beginning with consonant sequences, an infix can go inside or after the consonant sequence): *g-**um**-raduate* ~ *gr-**um**-aduate* 'graduated'
- infix *in* vs. prefix *ni*: *l-**in**-uto* ~ ***ni**-luto* 'to be cooked'
- location of reduplication in prefixed words: ***pa**-pag-lagy-an* ~ *pag-**la**-lagy-an* 'will place' (stem is *lagy*, from *lagay* 'location')

Variation in some of these phenomena has been investigated previously (Ross 1996 for partial reduplication; Rackowski 1999 for location of reduplication), sometimes using dictionary counts to obtain statistics (Zuraw 2002 for vowel height; Zuraw 2000 for nasal substitution). Corpus frequencies of the variants, however, or even basic word frequencies, have not previously been available.

As should be apparent from the examples given above, which are all in normal Tagalog spelling except for the hyphens added to show morpheme boundaries (hyphens are used in Tagalog, but not in the locations shown above), all of these phonological phenomena can be investigated in a text corpus. In most cases, modulo typing errors, we can be confident that the written form represents the writer's intended pronunciation, especially since spell-checking software that would change a writer's original spelling is not widely used for Tagalog, and there is little prescriptive pressure favoring one variant spelling over the other.[1] One area in which we should be cautious is partial reduplication, however: in a spelling such as *nag-fri-friendster*, it is plausible first that the writer might pronounce the stem in a nativized fashion despite preserving the English spelling (e.g., with [p] instead of [f][2]), and second that regardless of intended stem pronunciation, the reduplicant's spelling is merely an echo of the stem's spelling, and does not reflect the writer's pronunciation.

[1] Location of reduplicant is an exception: prescriptively, the reduplicant is adjacent to the root (Tania Azores-Gunter, p.c.).

[2] A Philippine social-networking website similar to friendster.com is jocularly named prendster.com.

Section 2 below describes how a written corpus of Tagalog was constructed from the web. Section 3 gives results from the corpus on tapping, and Section 4 concludes.

2 Construction of the corpus

Like most of the world's 6,000 or so languages, Tagalog is a language for which carefully constructed, tagged corpora (written or audio) do not exist. However, unlike most of the world's languages, Tagalog has a substantial web presence. As with all web-as-corpus endeavors, there is the drawback that the data will be messier, and there will be more input from non-native speakers than in, say, a newspaper-derived corpus. But in the case of some phenomena, such as infix location, a web corpus is actually preferable to a newspaper-derived corpus (if one existed): the range of loanwords found in formal Tagalog writing is narrower, favoring Spanish loans over English, than that found in the highly informal writing of blogs and web forums. From this informal writing we can obtain data on how the language's grammar is being extended to the novel phonological situations presented by a wide range of English loans.

A previous demonstration project (Ghani, Jones and Mladenić 2004) showed how a corpus of Tagalog can be created from the web by constructing queries designed to target Tagalog-language pages and exclude pages in other languages; the queries are created by using a small seed corpus to estimate word frequencies, and the frequencies are updated as the corpus grows. Kevin Scannell's *An Crúbadán* project (http://borel.slu.edu/crubadan/index.html), which seems to work in a similar fashion, includes a Tagalog language model. BootCaT (Baroni & Bernardini 2004), which is designed to create corpora and discover multi-word terms in specialized domains, such as psychiatry, works similarly, with the added twist that queries use words that are more frequent in the target domain than in a reference corpus. The method used here is similar, though cruder. No attempt is made to exclude pages written partly or even mostly in a language other than Tagalog; many blogs, for instance, are overwhelmingly in English but with occasional sprinklings of Tagalog, and I wanted to obtain these sprinklings, because they are rich in nonce affixed forms of loanwords.[3]

[3] I have not conducted any performance comparisons of different language-identification algorithms in pulling Taga-

In order to construct the corpus used here, first a smaller corpus of mainly Tagalog web pages, generously supplied by Rosie Jones (derived from Ghani, Jones and Mladenić 2004) was processed in order to yield estimated word frequencies for Tagalog.

Using these frequencies, a long list of queries composed of frequent words is automatically generated. Each term is at least 12 characters long, including spaces but not including apostrophes or other non-alphabetic characters. A word is chosen from among the most frequent 500 in the starter corpus, with a probability proportional to its log frequency. If this produces a 12-character string, the query is complete. Otherwise, another word is chosen using the same procedure and added to the string, until the threshold of 12 characters is reached. This threshold was selected in order to ensure queries long enough to be specifically Tagalog (e.g., not *sa ng*), but short enough to yield a large number of web hits. Some sample queries: *kami pangulo*, *+at salita oo*,[4] *lalo parang*, *noong akin aklat*. Although these queries are not treated as phrases, the order produced by the query-generator was preserved, because the topmost hits produced by, e.g., *lalo parang* and *parang lalo* are not the same. It is important to "toss the salad"[5] in this way, because the Google search engine that these queries are sent to allows only the top 1,000 results of a query to be viewed.

A program that sends these queries to Google (www.google.com), using the Google web APIs service, was written by Ivan Tam. This returns a maximum of 10,000 URLs (web addresses) per day, because a user's license key allows only 1,000 queries per day, and each query return only 10 results—to see more than the top 10 results for a given query, a new query must be sent, which counts against the day's 1,000. Typically, the number of URLs retrieved was about 5,000. This is because the number of times the program asks to see more results for a given query is determined by the estimated number of results initially reported by Google, but this is often an overestimate. For example, Google may estimate that there are 800 results, and the program will thus ask to see 80 pages of results (using up 80 of the day's queries), but perhaps only 621 results will be obtained. (The program gives the user the option of setting a maximum number of results to obtain per query; setting this number lower makes more efficient use of the day's query quota.)

Tam's program gives the option of taking using Google's option to return, out of any subset of results from one query that are highly similar, just one URL. That option was used here, but no further attempt was made to exclude highly similar results that come from different queries—obviously, this is an area where the procedure could be improved. The program also offers the option, which was used here, to create a separate query to search any crowded hosts (Google tends to show only two results from a single server, returning a "More results from ..." link; in the results returned by the Google Web APIs service, this translates into a non-blank value for `<hostName>`).

The day's URLs are compared against those retrieved so far, and the new ones are extracted. Another part of Tam's program then retrieves the full text of each new URL, although an existing program such as wget could also be used. Because the data of interest in this project are unigram and bigram frequencies, and irrelevant bigrams such as "`a href`" (a frequent bigram in html code) play no role, html stripping was not performed.

The resulting corpus currently has 98,607 pages and an estimated 20 million words of Tagalog (200 million "words" total, but examination of a sample finds that when html tags and non-Tagalog text are removed, about 10% remains). Word frequencies and certain bigram frequencies (e.g., the word+enclitic frequencies discussed below) are obtained from this corpus.

3 Tapping in the corpus

The phenomenon investigated most recently in the corpus is tapping. As mentioned above, Tagalog has a rule taking /d/ to the tap [ɾ] (spelled *r*) between vowels; tap rarely occurs non-intervocalically, except in loanwords (Spanish [ɾ] and [r], and English [ɹ] are usually adapted as [ɾ]). There are no opportunities for *d/r* alternation in affixes, but there are stems that begin or end in *d*, and if a vowel-final prefix or

log-language documents from the web, because this would require hand identification of their results (or of a large body of test documents). Qualitatively, however, the Ghani et al. approach does seem to suffer the same main problem as mine: a sizeable number of documents from Philippine languages other than Tagalog are retrieved.

[4] A "+" was added by hand to a few members of the top-500 list that Google would otherwise ignore because they are common function words in English or another major language. Quotations are placed around words with crucial punctuation, such as apostrophes in contractions.

[5] Thanks to Ivan Tam for this useful metaphor.

vowel-initial suffix is attached, the potential for tapping arises. Tapping has been reported to be variable at the prefix-stem boundary (*ma-**r**umi* 'dirty' vs. *ma-**d**ahon* 'leafy') but obligatory at the stem-suffix boundary (*laka**r**-an* 'to be walked on', from *laka**d*** 'walk') (Schachter and Otanes 1972). This is reminiscent of phenomena such as *s*-voicing in Northern Italian, which authors such as Nespor and Vogel (1986) and Peperkamp (1997) have analyzed as involving an asymmetry in how prefixes and suffixes relate to the prosodic word. For the sake of brevity, I will not review the Northern Italian facts here, but will apply an analysis similar to Peperkamp's to the Tagalog tapping case. (Peperkamp points out that prefix/suffix asymmetries always seem to be in this direction: prefixes are prosodically less integrated with stems than are suffixes.)

If we assume, as a first approximation, that a suffix is incorporated into the same prosodic word (p-word) as its stem, while a prefix adjoins to the stem to form a higher p-word, and we further assume that tapping applies only to a vowel-*d*-vowel stretch that is not interrupted by a p-word boundary, then we would predict that tapping occurs at the stem-suffix boundary but not at the prefix-stem boundary:

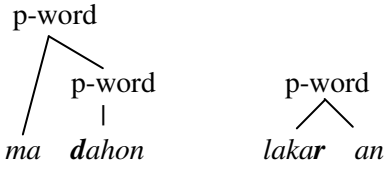

Figure 1. Prosodic structure of prefixed word without tapping vs. suffixed word.

Loosely following Peperkamp, I will assume that this prosodification is derived by a constraint requiring the left edge of any accessed lexical unit (see below) to project the left edge of a p-word. In Optimality Theory terms (Prince & Smolensky 1993/2004), the symmetrical constraint requiring the right edge of an accessed lexical unit to project any prosodic edge is ranked lower (specifically, below an anti-recursion constraint requiring every p-word node to immediately dominate a foot).

3.1 Tapping at the prefix-stem boundary

How can we explain *ma-**r**umi*, where tapping does occur at the prefix-stem boundary? In the Northern Italian case, Baroni (2001) found that application of the *s*-voicing rule at the prefix-stem boundary in a reading task was negatively correlated with semantic transparency as determined by a rating task. Baroni's interpretation is that forms with voicing (which tend to be semantically opaque) are treated as morphologically simple. I will follow Baroni loosely in assuming that words like *marumi* are accessed as a single lexical unit (without taking a position on whether that lexical entry contains information about morpheme boundaries). If *marumi* is accessed as a lexical unit—rather than indirectly via *ma-* and *dumi*—then the constraint mentioned above requires only the left edge of the whole word to project a p-word boundary, and the structure is as in Figure 2. Because no p-word boundary interrupts the vowel-*d*-vowel sequence, tapping applies.

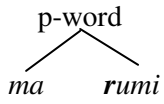

Figure 2. Prosodic structure of prefixed word with tapping.

The corpus does not directly yield judgments of semantic transparency, of course—though indirect measures using the similarity of contexts in which the derived word and its base occur could be examined in future work—but it does yield a statistic that Hay (2003) has argued is closely related to the degree to which a morphologically complex word is treated as a single unit vs. compositionally: the ratio of base frequency to derived-word frequency. Hay argues, based on a series of experiments on English, that when a derived word is more frequent than its morphological base (e.g., English *illegible* vs. *legible*), it is more likely to be accessed through a direct route during processing (direct access to *illegible* rather than access via *in-* and *legible*), and thus more likely to be treated as a single unit phonologically, and more likely to develop independent semantics. The prediction that can be tested in the Tagalog corpus is this: prefixed words that are more frequent than their unprefixed bases are more likely to undergo tapping than prefixed words that are less frequent than their unprefixed bases.

To minimize hand-checking of items, the corpus was searched only for the 592 orthographically distinct prefixed *d*-stem words that appear in a dictionary of Tagalog (English 1986). These words were extracted from the dictionary and put into electronic form by Nikki Foster. The frequency of each word's tapped and untapped form

were retrieved from the corpus (e.g., for the dictionary's *i-**d**ipa*, both *idipa* and *iripa*'s frequencies were obtained). Dictionary-listed variants were searched, and certain punctuation was allowed. "Linkers" were also allowed (these are clitics that can become, orthographically, part of the preceding word). The frequency of each word's root, as listed in the dictionary, was also retrieved. (In the case of words with multiple affixes, it is unclear what the immediate morphological predecessor is, so for the sake of simplicity the root, rather than some intermediate form, was used.)

The histograms below show how many prefixed words display each range of tapping rates in the corpus, from 0 (always *d*) to 1 (always *r*). They demonstrate the predicted influence of derived/base frequency ratio on tapping rate: when the prefixed word is more frequent than its root (Figure 3), a high rate of tapping predominates (strongly), whereas when the root is more frequent than the prefixed word (Figure 4), a low rate of tapping predominates (weakly):

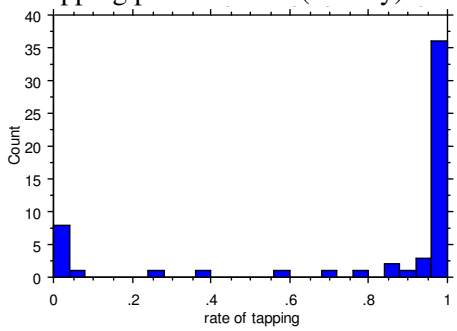

Figure 3. Distribution of tapping rate in prefixed words that are *more* frequent than their bases.

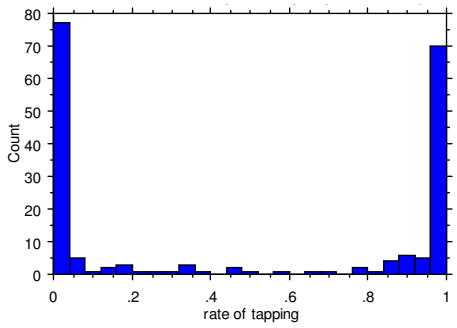

Figure 4. Distribution of tapping rate in prefixed words that are *less* frequent than their bases.

Interestingly, in both cases the rates of tapping cluster near 0 and 1—intermediate rates are relatively rare. The data above are limited to words with a corpus frequency of at least 10, so that each word had a fair chance of displaying an intermediate rate of tapping if that were its true behavior. This suggests that the great majority of prefixed words in Tagalog are lexicalized as either undergoing or not undergoing tapping (or, depending on what form lexical entries in fact take, as having one prosodic structure or the other). This is rather different from the Northern Italian situation discovered by Baroni, where many words robustly vary, even within a single speaker.

Words with a corpus frequency of less than 10, which are almost all less frequent than their bases, show a preference of non-tapping, as expected:

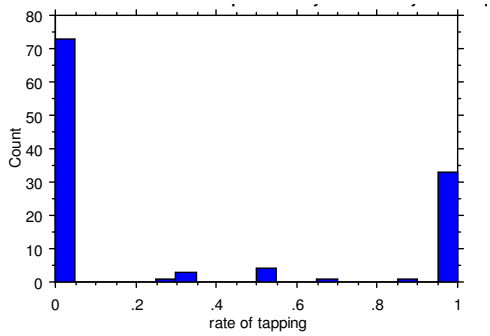

Figure 5. Distribution of tapping rate in prefixed words with corpus frequency < 10 (nearly all are less frequent than their bases).

Hay argues that it is relative frequency of a derived word and its base, not raw frequency of the derived word, that models of lexical access predict to have an effect on word decomposability. In the present case, raw frequency does also have a strong effect on whether a prefixed word belongs to the tapping or non-tapping categories, but raw and relative frequency are themselves highly correlated. In order to verify that relative frequency has an effect independent of raw frequency, the prefixed words were divided into 28 categories according to the log of their raw frequency (0 to <0.1, 0.3 to <0.4, 0.4 to <0.5, etc.). Within each category, the percentage of words *less* frequent than their bases that undergo tapping >95% of the time and the percentage of words *more* frequent than their bases that undergo tapping >95% of the time were calculated. The prediction is that the second percentage should be higher—that is, words matched for raw frequency should be more likely to undergo tapping if they are more frequent than their bases—and this was borne out in a Wilcoxon signed-

rank test (p<.05). The contribution of raw frequency remains to be further explored.

3.2 Tapping at the stem-suffix boundary

Tapping was examined in a similar fashion at the stem-suffix boundary. From English's (1986) dictionary, 160 native-etymology roots that end in *d* were extracted, and the corpus was searched for any suffixed forms of these roots (with or without additional prefixes and infixes). As expected from Schachter and Otanes's (1972) description, tapping is indeed nearly obligatory at the stem-suffix boundary, as shown in Figure 6 (which again shows only words with corpus frequency of at least 10):

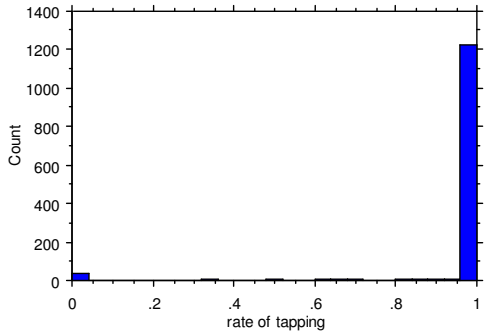

Figure 6. Distribution of tapping rate in suffixed words.

Because of the relative ease of searching for suffixed forms (there are only two productive native suffixes in Tagalog, *-in* and *-an*, so a simple regular expression can find all the suffixed forms of any root), the counts here are much higher than in the prefix-stem case—compare the scales of the vertical axes in the histograms—and we can look more closely at the 124 words—a minority so small it is largely invisible in Figure 6—that do not uniformly undergo tapping at the stem-suffix boundary. Rate of tapping among these 124 words turns out to be weakly but significantly correlated with the log ratio of suffix-word frequency to root frequency (Spearman's rho=.534, p<.001), as predicted by Hay's view of phonological integration.

There are multiple possible interpretations for this result under the prosodic account given above. Perhaps stem and suffix do always form a single p-word, but paradigm-uniformity effects (e.g., Steriade 2000) can, if sufficiently strong, block tapping even within a p-word. Or, perhaps the requirement that a suffix be integrated into the prosodic word can itself be overridden, occasionally, by frequency effects demanding a compositional treatment of an affixed word that is less frequent than its base.[6] It is also possible that all the "nontapping" here represents typographical errors, but that there is a frequency effect on errors such that the more frequent a base relative to the word it is nested inside, the more likely that the base's spelling is preserved.

3.3 Tapping at the stem-stem boundary

The prosodic system assumed above (with some constraints not mentioned there), allows a combination of two stems to have either of the prosodic structures shown in Figure 7, with the choice depending on whether the combination is accessed as a single lexical unit. But in either case, a p-word boundary separates the two stems, and thus tapping is not expected on either side of the stem-stem boundary.

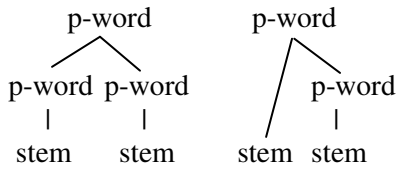

Figure 7. Two possible prosodic structures for compound or two-syllable reduplication.

There are two places where a stem+stem combination could arise in Tagalog. One is in compounds, such as *basag-ulo* 'fight' (lit. *breaking-head*), where each member bears a separate stress. If we assume, following most previous work on the p-word, that dominating a stressed syllable is a necessary feature of a p-word (though not sufficient, since a single p-word may contain multiple stresses), this is consistent with a p-word+p-word prosodic structure. Lacking a list of compounds, however, I found it impractical to search for compounds in the corpus (though this is a project for the future).

The second place where stem-stem boundaries arguably arise is in two-syllable reduplication, which occurs in a variety of morphological constructions, including reduplication by itself: e.g. *pa-balik-balik* 'recurrent', from *balik* 'return'. In these reduplications, each copy bears a stress. We would therefore expect that tapping should not occur at the boundary between the two redu-

[6] In Hay's view, relative frequency is not epiphenomenal, but rather determines the mode of lexical access (direct or indirect route) and thus a word's behavior. It is also possible, of course, that relative frequency is only the symptom of some underlying property of words, or that there is feedback between frequency and the properties that influence it.

plicants. This is indeed what is found, as shown in the histogram below, though the data come mostly from stem-initial *d* cases (e.g., *dagli-**d**agli* 'right away'); there were only 5 stem-final *d* cases that met the frequency threshold (e.g., *aga**d**-agad* 'immediately'):[7]

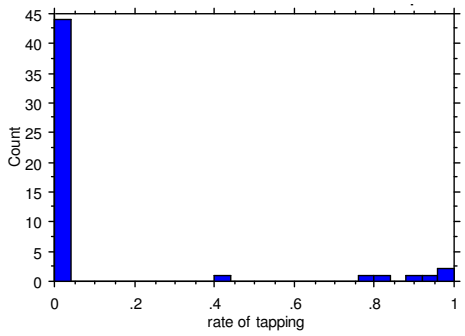

Figure 8. Distribution of tapping rate at reduplicant-reduplicant boundary (two-syllable reduplication).

The lack of tapping is unlikely to be a reduplicative identity effect (Wilbur 1973, McCarthy and Prince 1995), because tapping is blocked even when the other copy of the same consonant does undergo tapping because of an adjacent prefix or suffix (*ka-aga**d**-aga**r**-an*, *ka-**r**aga-**d**aga-_n* [glosses unknown—English's dictionary contains both roots but not these derivatives of them]).

The lack of tapping is also probably not due to the reduplicated forms' low frequency: most are indeed less frequent than their bases, but it was seen above that prefixed words that are less frequent than their bases undergo tapping almost as often as not.

3.4 Tapping in clitics

There are two enclitics in Tagalog that begin with /d/: *din* 'also' and *daw* '(reported speech)'. Each has a tap-initial allomorph (*rin*, *raw*). There is reported to be variation between the two allomorphs even after consonant-final words (Schachter and Otanes 1972). So far, I have examined in the corpus only *din/rin* after vowel-final words.

All bigrams whose second word is *din* or *rin* were extracted from the corpus. Variation was indeed found, but unlike in the prefix+stem case, where the variation was highly polarized—with most words having one strongly dominant behavior—in the word+clitic case the variation is continuous (again, only bigrams with a corpus frequency of at least 10 are shown):

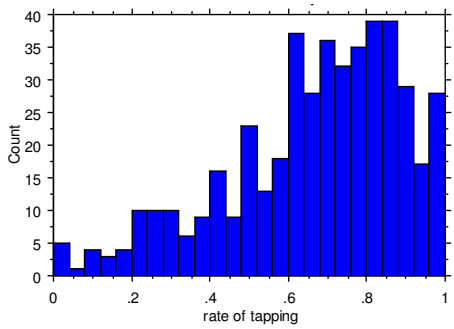

Figure 9. Distribution of tapping rate at word-clitic boundary.

One interpretation is that most word+clitic combinations are not lexicalized, and their tapping behavior is determined on the fly. The correlation between the log ratio of bigram to base word frequency and the rate of tapping, though very weak, is significant (Spearman's rho=.197, $p<.0001$). If we look at enclitic+din/rin combinations (where the first enclitic ends in a vowel, as in *... pa rin* '... still also'), which display similarly gradient variation, the correlation is stronger, though p is larger because there are fewer data points (Spearman's rho=.527, $p<.05$).

4 Conclusion

This paper has presented one case study, on Tagalog tapping, of phonological research using a written, web-derived corpus. Several aspects of the investigation depended crucially on the web-as-corpus method. Because of economic constraints, the only realistic way to assemble a large corpus of a language like Tagalog is currently by taking text from the web. And only a large corpus makes it possible to ask questions such as "how does the frequency ratio of a derived word to its base affect the application of a phonological rule?" The two different patterns of variation—polarized in the stem+prefix case, continuous in the word+enclitic case—would have been very difficult to discover without corpus data.

This Tagalog corpus has already been used to investigate infixation in loans that begin with consonant clusters (Zuraw 2005). There, as mentioned in Section 2, the web-based nature of the

[7] The interpretation of stem-final *d* cases is complicated by the fact that p-words spelled with an initial vowel are usually actually glottal-stop initial. Thus, *aga**d**-agad* can be pronounced with a glottal stop (*aga**d**-[ʔ]agad*), so that the medial *d* is not truly intervocalic.

corpus was of more than practical importance, because a large quantity of highly informal writing—unlikely to be found in a traditionally constructed written corpus—was needed.

The corpus is also being used in ongoing work on nasal substitution, and will be used in the future to investigate the other phenomena listed in Section 1. The corpus will also continue to grow; there seems to be little danger of running out of Tagalog-language web space to search in the foreseeable future.

Acknowledgement

Thanks to research assistant Ivan Tam for programming that made this project possible, and to research assistant Nikki Foster for data entry. For valuable discussion about tapping, thanks to Colin Wilson, Bruce Hayes, and participants the UCLA phonology seminar. Thanks also to two anonymous reviewers for several ideas that have been incorporated into the paper.

References

Baroni, Marco (2001). The representation of prefixed forms in the Italian lexicon: Evidence from the distribution of intervocalic [s] and [z] in northern Italian. In Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology* 1999, Springer, Dordrecht: 121-152.

Baroni, Marco and S. Bernardini (2004). BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*.

English, Leo (1986). *Tagalog-English Dictionary*. Congregation of the Most Holy Redeemer, Manila. Distributed by (Philippine) National Book Store.

Ghani, Rayid, Rosie Jones & Dunja Mladenić (2004). Building minority language corpora by learning to generate Web search queries. *Knowledge and Information Systems* 7: 56-83.

Hay, Jennifer (2003). *Causes and Consequences of Word Structure*. Routledge, New York and London.

McCarthy, John & Alan Prince (1995). Faithfulness and reduplicative identity. *Papers in Optimality Theory, UMass Occasional Papers in Linguistics* 18: 249-348

Nespor, Marina and Irene Vogel (1986). *Prosodic Phonology*. Foris, Dordrecht.

Peperkamp, Sharon (1997). *Prosodic Words*. Holland Academic Graphics, The Hague.

Prince, Alan and Paul Smolensky (1993/2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Rackowski, Andrea (1999). Morphological optionality in Tagalog aspectual reduplication. *Papers on Morphology and Syntax, Cycle Two, MIT Working Papers in Linguistics* 34: 107-136.

Ross, Kie (1996). Floating phonotactics: infixation and reduplication in Tagalog loanwords. UCLA M.A. thesis.

Schachter, Paul and Fe Otanes (1972) *Tagalog Reference Grammar*. University of California Press, Berkeley.

Steriade, Donca (2000). Paradigm Uniformity and the phonetics/phonology boundary. In Janet Pierrehumbert and Michael Broe (eds.), *Papers in Laboratory Phonology* vol. 6, Cambridge University Press, Cambridge.

Wilbur, Ronnie Bring (1973). *The Phonology of Reduplication.* Indiana University Linguistics Club, Bloomington.

Zuraw, Kie (2000). Patterned exceptions in phonology. UCLA Ph.D. dissertation.

Zuraw, Kie (2002). Aggressive reduplication. *Phonology* 19: 395-439.

Zuraw, Kie (2005). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog. Manuscript, UCLA.