Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20). 2008. Volume 1.
Edited by Marjorie K.M. Chan and Hana Kang. Columbus, Ohio: The Ohio State University. Pages 109-122.

The Spotty-Data Problem in Phonology[*]

San Duanmu
University of Michigan

A common goal of a phonologist is to describe the phonology of a language. The job would be easier if the boundaries of phonology are clear, but they are often not. I discuss some examples, especially what I call the spotty-data problem, which refers to the fact that we often do not have enough data to figure out certain rules or constraints in a language. I also offer an explanation why phonological data are so spotty and suggest that sometimes universal generalizations might be easier to determine than language-specific ones.

1. What is the phonology of a language?

Phonologists often work to describe the phonology of a language. But what exactly is the phonology of a language? According to Halle (1962), it is a set of rules. This conception follows from Chomsky's (1957) proposal that a grammar is a set of rules that define a set of possible and impossible linguistic structures. The set of rules is limited, but the set of structures defined by them can be infinite, because some rules can be 'recursive'. For example, the recursive rules in (1) can generate infinitely long sentences and the recursive rules in (2) can generate infinitely long words.

(1) S → NP VP
VP → V S
VP → V NP

An infinitely long but grammatical sentence:
I know you know I know you know…that you have a dog.

[*] This is a revised version of a longer paper to appear in Interfaces in Chinese Phonology: Festschrift in Honor of Matthew Chen, edited by Yuchau Hsiao, Hui-Chuan Hsu, and Lian-Hee Wee. Apart from the acknowledgments made in the longer paper, I would like to thank the audiences at NACCL-20 and Institute of Linguistics, Chinese Academy of Social Sciences for their comments.

(2) N → A-*ness*
A → N-*less*

Grammatical words, which can be infinitely long:
red-ness, red-ness-less, red-ness-less-ness, red-ness-less-ness-less,…

The sentence or words in (1) and (2) may have never been used before, but they are nevertheless grammatical, because they can be derived from the rules of English.

Chomsky's proposal outlines generative grammar, in which the set of rules are well defined, so are the set of structures they generate, whether the structures are limited in size or infinitely long, and whether the structures are familiar or new. In other words, if we can figure out the grammar of a language, we can predict or generate all and only correct structures in the language.

Halle (1962) offers a similar view of what phonology is. According to him, the phonology of a language is a set of rules that can generate all and only correct phonological forms of a language. In addition, because the speaker of a language has the knowledge of the grammar, she or he has the intuition to judge whether a structure is or is not good in the language, whether the structure has been used before or not. In particular, speaker intuition can be used to judge whether a sound sequence is well formed, even if it is not a real word, as shown in (3).

(3) Speaker judgment on potential words in English (Halle 1962)
a. Good (possible words): [bɪk], [θod], and [nɪs]
b. Bad (impossible words): [tsaim], [gnait], and [vnɪg]

None of the words in (3) are (or were) real English words but those in (3a) are possible words while those in (3b) are not. This conception of what phonology (or grammar) is has lead to extensive rule-based research for the next 20-30 years, such as Chomsky and Halle (1968) on English phonology and Cheng (1973) and Lin (1989) on Chinese phonology.

Chomsky (1981) replaces grammatical rules with principles, parameters, and lexical specifications. Similarly, Prince and Smolensky (1993) argue that rules should be replaced with ranked constraints. However, the essence of the generative proposal remains the same: the grammar and the structures it generates are well defined, so is the knowledge or intuition of the speaker.

2. Gradient intuition

However, many studies have found that the intuition of native speakers is not always clear. For example, Frisch et al (2000), Myers and Tsay (2005), and Zhang (2007) have shown that speaker judgment on possible words is not clear cut but gradient. Similarly, many studies have noted that consistent judgment on syllable boundaries can be hard to obtain (e.g. Gimson 1970, Treiman and Danis1988, Giegerich 1992, Hammond 1999, Steriade 1999, Blevins 2003), and this has lead to different analyses in some cases. An example is shown in (4), where a dot indicates a syllable boundary and [t̲] is an ambisyllabic [t] (a single [t] that belongs to both the first and the second syllable).

(4) Syllable boundary in *city*

Pulgram (1970), Kahn (1976):	*cit̲y*	(ambisyllabic [t])
Selkirk (1982):	*cit.y*	
Halle and Vergnaud (1987):	*ci.ty*	
Burzio (1994):	*cit.ty*	(geminate [tt])

The lack of speaker intuition on syllable structure may be taken as evidence that there are no syllables, a position held by Chomsky and Halle (1968). It is also possible thought that not everything is intuitively obvious. For example, we are not exactly aware of how we see colors, how we digest food, or how we walk. Similarly, if we rely on intuition alone, we would wrongly conclude that the earth is flat. Indeed, although many phonologists agree that words are made of consonants and vowels and that consonants and vowels are in turn made of distinctive features, the assumptions are far from obvious to the average person, or even to some linguists (such as Ladefoged 2001). The lack of intuitive judgment on linguistic structures means that linguists have to work harder in figuring out patterns of grammar.

Frisch et al (2000) have shown that speaker judgment on possible words (non-words) in English is based on frequency. A non-word whose parts occur frequently in existing words, such as those in (4), is likely to be judged better than one whose parts occur infrequently, such as those in (5).

(5) Non-words whose parts occur frequently in English words
[midət], [kinəp], [henət], …

(6) Non-words whose parts occur infrequently in English words
[zoɪʒəʃ], [jʊgoɪl], [vɔθəʃ], …

Frisch et al (2000) can also explain why speaker judgment is sometimes quite clear, as noted by Halle (1962): [bɪk], [θod], and [nɪs] are judged to be possible words because their parts occur frequently in English words, whereas [tsaim], [gnait], and [vnɪg] are judged to be impossible words because [ts-, gn-, vn-] never occur as onsets in English.

Still, an important question remains: How often are boundaries of phonology vague? For example, do speakers have clear judgment 95% of the times and uncertain judgment just 5% of the times? Or is the extent of uncertainty far greater than that?

3. The spotty-data problem

I shall show that the uncertainty in phonological analysis is far more extensive than previously conceived. I shall call it the spotty-data problem (Duanmu 2008), which refers to the fact that there are often not enough data for making reliable generalizations, even if we examine the entire lexicon of a language. To begin, let us consider the ratios between possible words and actual words in English. First, consider the number of possible CVC syllables in American English, shown in (7), where V is a short (lax) vowel.

(7) CVC syllables in American English

Initial C:	23	[p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, ʒ, h, tr, dr, ʧ, ʤ, m, n, l, r][1]
Lax V:	5	[ɪ, ʊ, ɛ, ʌ, æ]
Final C:	21	[p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, ʒ, ʧ, ʤ, m, n, ŋ, l, r]
CVC:	2,415	

American English has 24 consonants, including the affricates [tr, dr, ʧ, ʤ]. Of these 23 can occur in the onset (excluding [ŋ]) and 21 can occur in the coda (excluding [h, tr, dr]. American English also has 5 short (lax) vowels. This gives 23 x 5 x 21 = 2,415 possible CVC syllables. However, the actual number of occurring CVC syllables is a lot smaller, as seen in (8), based on the CELEX lexicon (Baayen et al 1993).

(8) The spotty-data problem in English

Word form	Possible	Used	% used
CVC	2,415	615	25.5%
CVCCVC	5,832,225	6,000	0.1%

Excluding affixes and homophones, English has about 3,000 uninflected monosyllabic words, which include CVVC (842), CVC (615), CCVVC (453), CCVC (326), etc. The second most frequent type, CVC, includes 615 syllables. This means that just one fourth of all possible CVC words are used. In dialects that have more short vowels, the percentage of occurring syllables could be even lower. If we consider disyllabic words, the percentage of occurring syllables becomes diminishingly small. For example, if any two CVC syllables can form a disyllabic word, there are about 6,000,000

[1] Whether [tr, dr] are single sounds (affricates) or clusters of two each does not affect the point being made here. Phonetically, [tr, dr] are affricates, as noted by many phoneticians, such as Jones (1950), Abercrombie (1967), Gimson (1970), and Wells (1990). Phonologically, one might ask whether stop + approximant can create an affricate. The answer is yes, such as [t] + [j] → [ʧ] in *get you* and [d] + [j] → [ʤ] in *did you*. One might also ask that, if we treat [tr, dr] as affricates, would we increase the English phoneme inventory by two? The answer is no; all we need to say is that when the phonemes [t, r] or [d, r] occur in the same onset, they form an affricate.

possible disyllabic words, yet English only uses around 6,000 uninflected disyllabic words. This means that just 0.1% of all possible disyllabic words are used.

Because there are so few occurring forms, it is often hard to determine what the general pattern is, or what rules or constraints one should propose. As an example, let us take a closer look at monosyllables in English. English has at least 59 productive onsets, shown in (9)-(12). The distinction between productive and unproductive onsets is not a technical one. Unproductive onsets mostly include those that only occur with one vowel; for example, Cj and CCj only occur with the vowel [u]. Should we include all onsets in the calculation, the spotty-data problem would be even more serious.

(9) Occurring productive onsets in English (59 in all)
C onsets (22 in all)
CC onsets (30 in all)
CCC onsets (6 in all):
Lack of an onset (1 in all)

(10) Productive C onsets (22 in all):
p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, h, tr, dr, ʧ, ʤ, m, n, l, r
Unproductive C onsets (1 in all):
ʒ

(11) Productive CC onsets (30 in all):
bl (*black*), br (*bring*), dr (*dry*), dw (*dwell*), ʃr (*shrink*), ʃw (*schwa*), fl (*fly*), fr (*fry*), gl (*glad*), gr (*green*), gw (*penguin*), kl (*class*), kr (*cry*), kw (*quick*), pl (*plot*), pr (*price*), sl (*sleep*), sw (*swim*), tr (*try*), tw (*twin*), θr (*three*), θw (*thwart*), st (*stop*), sp (*spot*), sk (*sky*), sn (*snake*), sm (*smack*), sf (*sphere*), ʃm (*schmaltz*), ʃn (*schnitzel*)
Unproductive CC onsets (36 in all):
bj (*beauty*), dj (*duty*), fj (*few*), gj (*argue*), hj (*huge*), kj (*cute*), lj (*volume*), mj (*music*), mw (*moiré*), nj (*news*), nw (*peignoir*), pj (*pure*), pw (*puissance*), sj (*suit*), sr (*Sri Lanka*), tj (*tube*), vj (*view*), vw (*reservoir*), zl (*zloty*), zj (*presume*), ʒw (*bourgeois*), θj (*enthuse*), km (*Khmer*), kn (*Knesset*), kv (*kvass*), sv (*svelte*)

(12) Productive CCC onsets (6 in all):
str (*string*), skr (*screen*), skw (*square*), spr (*spring*), spl (*splash*), skl (*sclerosis*)
Unproductive CCC onsets (4 in all):
stj (*studio*), skj (*skew*), spj (*spew*), tsw (*Tswana*)

Given 59 productive onsets, we expect there to be at least 59 different monosyllables with the rhyme [ɪl] (the most frequent VC rhyme). However, only 29 of them occur in the CELEX lexicon. The 30 non-occurring ones are shown in (13).

(13) Non-occurring monosyllables with the rhyme [ɪl]
vɪl, θɪl, ðɪl, zɪl, ʃɪl, lɪl, jɪl; blɪl, dwɪl, ʃmɪl, ʃnɪl, ʃwɪl, flɪl, glɪl, gwɪl, klɪl, krɪl, plɪl, prɪl, sfɪl, slɪl, smɪl, snɪl, θwɪl; strɪl, skrɪl, skwɪl, sprɪl, splɪl, sklɪl

A few of the syllables may be used in words that CELEX failed to collect, such as *shill* and *krill*. It has also been proposed that there is a constraint against C+[lɪl] (Clements and Keyser 1983: 21, Davis 1988: 25), or against [lɪl] in general (Pierrehumbert 1994: 186), although there is a word *lilt*. Still, there are many others left, which seem to be accidental gaps, because they do not seem to violate any obvious phonological requirement.

Besides accidental gaps, there are forms that seem to be outliers (or exceptions)—those that do not seem to fit the patterns of other syllables. For example, [ts] is rarely used as an onset in English, but it occurs in *Tswana* [tswa:][na] and *scherzo* [skeɚ][tso]. Similarly, [s] does not occur with a fricative in word-initial position except in *svelte*, *sforzano*, *sphagnum*, *spheroid*, *sphincter*, *sphinx*, and *sphere*. Most of these words can probably be labeled as foreign or uncommon, although it is hard to rule out *sphere* this way. In Chinese there are outliers, too. For example, Cantonese generally disallows two labial sounds in a syllable, but it has the word [pʌm] 'pump'. Similarly, in Standard Chinese a palatal onset usually does not go with a diphthong that ends in [i], but then there is a marginal word [jai] 'cliff', which most people pronounce as [ja].[2]

If we always know what accidental gaps are and what outliers are, it can still be reasonably easy to figure out the rules in a grammar. But the problem is that it is not always clear whether an occurring form is an outlier or a good word, nor is it easy to decide whether a non-occurring form is a potential word or simply ungrammatical. Our decisions on such cases would lead to fairly different versions of English phonology. For example, (14) and (15) show two ways to treat *Tswana* and *scherzo*.

(14) Decision:
Tswana and *scherzo* are outliers (not good words) in English.[3]

[2] Two reviewers point out that few speakers use [jai] any more, but [ja] instead. The point remains the same though: before [jai] dropped out of use, it was the only word of the form [jVi], where V is any vowel. Was it an outlier then, or was it well-formed but simply infrequent?

[3] A review points out that *Tswana* and *scherzo* are clearly borrowed foreign words. If we exclude them, English has no onset [ts]. How can the lack of onset [ts] in English be explained? The answer I suggest is that a language does not need to use every possible onset. In fact, this is

Generalizations:

English words and syllables cannot start with [ts].

[tsæt, tsɪl, …] are impossible (ungrammatical) words in English.

(15) Decision:

Tswana and *scherzo* are good words (not outliers) in English.

Generalizations:

English words and syllables can start with [ts].

[tsæt, tsɪl, …] are potential words in English.

Similarly, the decision on words like *sphere* can lead to two analyses, shown in (16) and (17).

(16) Decision:

sforzano, *sphagnum*, *spheroid*, *sphincter*, *sphinx*, and *sphere* are outliers (not good words) in English.

Generalizations:

English words and syllables cannot start with [sf].

[sfɪt, sfain, …] are impossible (ungrammatical) words in English.

(17) Decision:

sforzano, *sphagnum*, *spheroid*, *sphincter*, *sphinx*, and *sphere* are good words (not outliers) in English.

Generalizations:

English words and syllables can start with [sf].

[sfɪt, sfain, …] are potential words in English.

If we treat [sf] as an outlier, we expect [sfɪt] *sfit* and [sfain] *sfine* to be ungrammatical in English (they seem to be as marginal as *Tswana* or *sforzano*). On the other hand, if [sf] is not an outlier, we expect [sfɪt] *sfit* and [sfain] *sfine* to be accidental gaps or potential words in English.

As another example, consider occurring and non-occurring monosyllables with VC rhymes in English again, shown in (18).

expected from the spotty-data problem. It is worth noting that the lack of onset [ts] in English is not because [ts] is ill-formed in any way: German uses it without any problem, and English could adopt it any time when words like *Tswana* and *scherzo* become part of the daily vocabulary, or the vocabulary of English learning children.

(18) English monosyllables with VC rhymes

Productive onsets:	59
Occurring VC rhymes:	101
Possible monosyllables with VC rhymes:	5,959
Occurring monosyllables with VC rhymes:	1,069

As seen earlier, English has 59 productive onsets. In addition, English has 101 occurring VC rhymes. Therefore, there are 5,959 possible monosyllables with VC rhymes. However, only 1,069 occur in the CELEX lexicon. The 101 VC rhymes are shown in (19). The number of occurring onsets for a rhyme is in parentheses and the CELEX [O] can be [ɒ] or [ɑ] in American English.

(19) VC rhymes and the number of onsets they occur with in monosyllables
[ɪl] (29), [ɪp] (26), [æk] (25), [ɪt] (25), [Ot] (25), [æt] (24), [ɪk] (23), [Op] (23), [æg] (22), [æʃ] (22), [æp] (21), [æm] (20), [ʌm] (20), [ɪn] (19), [Ok] (19), [ʌg] (19), [æd] (18), [æn] (18), [ɛd] (18), [Ob] (18), [Od] (18), [ʌf] (18), [ɛl] (17), [ɛn] (17), [ɛt] (17), [ɪŋ] (17), [Og] (16), [ʌb] (16), [ʌt] (16), [æb] (15), [ɪm] (15), [ɪg] (14), [ɪʧ] (14), [ʌk] (14), [ɛs] (13), [Os] (13), [Oʃ] (13), [ʌn] (13), [ʌʃ] (13), [æŋ] (12), [ɛk] (12), [ɪf] (12), [ɪb] (11), [ɪz] (11), [ʌd] (11), [æʧ] (10), [ɛʤ] (10), [ɪd] (10), [On] (10), [Oŋ] (10), [ʌs] (10), [ʌʤ] (9), [ʌŋ] (9), [ɛg] (8), [ɛʧ] (8), [ɪs] (8), [Of] (8), [Oʧ] (8), [ʊk] (8), [ʌl] (8), [æs] (7), [ɛf] (7), [ɛm] (7), [Ol] (7), [Om] (7), [ʌʧ] (7), [ɛp] (6), [ɪʃ] (6), [Oθ] (6), [ʊd] (6), [Oʤ] (5), [ʌp] (5), [ʌv] (5), [æl] (4), [ɛb] (4), [ɛʃ] (4), [ɪʤ] (4), [ɪθ] (4), [ɪv] (4), [ʊl] (4), [ʊʃ] (4), [ʊt] (3), [ʌz] (3), [æʤ] (2), [æf] (2), [æv] (2), [æz] (2), [ɛθ] (2), [ʊʧ] (2), [æθ] (1), [əl] (1), [əm] (1), [əs] (1), [əv] (1), [ɛv] (1), [ɛz] (1), [ɪð] (1), [Ov] (1), [Oz] (1), [ʊf] (1), [ʊs] (1)

The non-occurring monosyllables with high-frequency rhymes are probably accidental gaps, although even the most frequent rhymes hardly occur with half of the onsets. A harder question is what to do with low frequency rhymes, such as those that only occur with one onset each. For example, [ɪð] occurs in just one monosyllable (alternative pronunciations excluded), which is *with*. Is *with* an outlier (exception)? The answer again leads to two analyses, shown in (20) and (21).

(20) Decision:
with is an outlier (not good word) in English.

Generalizations:

[ɪð] is not a possible rhyme in English.

[mɪð, nɪð, tɪð,…] are impossible words in English.

(21) Decision:

with is a good word (not an outlier) in English.

Generalizations:

[ɪð] is a good rhyme in English.

[mɪð, nɪð, tɪð,…] are potential words in English.

If *with* is an outlier (phonologically bad but still used as an exception) in English, non-words such as [mɪð, nɪð, tɪð, kɪð, …] would be ungrammatical. If *with* is phonologically good (not an outlier) in English, the same non-words would be accidental gaps and potential words.

Next consider Cantonese. Yip (1988: 82) suggests that Cantonese Chinese has a restriction against syllables that have two labial consonants, one in the onset and one in the coda, such as [pim] and [map], but she notes a few exceptions, such as [pʌm] 'pump'. Should we say that Cantonese disallows syllables with two labial consonants, or should we say that Cantonese is in principle open to their use, but happens not to have used any (or many) such words? The two options are shown in (22) and (23).

(22) Decision:

[pʌm] 'pump' is an outlier (not good word) in Cantonese.

Generalizations:

Cantonese allows no syllable with a labial onset and a labial coda.
[pau, mau, …] are impossible words in Cantonese.

(23) Decision:

[pʌm] 'pump' is a good word (not an outlier) in Cantonese.

Generalizations:

Cantonese allows syllables to have a labial onset and a labial coda
[pau, mau, …] are potential words in Cantonese.

Indeed, one might also want to ask: why should [pʌm] 'pump' be an outlier in Cantonese, if *map* and *Pam* are perfect syllables in English?

Next consider Standard Chinese. We mentioned earlier that Standard Chinese generally lacks syllables that have a palatal or front glide in the onset and a front high vowel in the coda, but there is one word [jai] 'cliff', which many people pronounce as

[ja]. Is [jai] an outlier so that Standard Chinese has a restriction against two front high vowels (Lin 1989, Duanmu 2000), or is [jai] a good syllable and there is no such restriction? Also, why should [jai] 'cliff' be an outlier in Standard Chinese, if [tɕai] 'release' occurs in other Mandarin dialects, such as Chengdu? Consider another case in Standard Chinese, where the medial glide cannot be [ɥ] if the initial C is a sonorant, such as [nɥa], [nɥan], [nɥauŋ], [lɥa], [lɥan], and [lɥauŋ], but there are two exceptions, [nɥe] 'mistreat' and [lɥe] 'abbreviate'. Should these syllables be outliers, or should we say that most syllables with a medial [ɥ] happen to be accidental gaps?

A reviewer suggests that 'if we can foresee the future of a language, it won't be difficult to make a decision' on whether a form is an outlier or not. For example, if we look at Standard Chinese alone, we might be unsure whether [nɥe] and [lɥe] are outliers, but if we look at other Chinese dialects, many of which lack any sonorant + [ɥ] combination, or if we look at historical trends, where sonorant + [ɥ] combinations seem to be dropping out, then we might conclude that [nɥe] and [lɥe] are outliers. It is true that sometimes we can tell whether a form is dropping out of a language, such as [jai] in Standard Chinese. However, to 'foresee the future of a language' in general is far from easy, and I am not aware of any serious proposal in this regard. In addition, outliers and accidental gaps are synchronic notions, not diachronic ones. For example, if CCV syllables are evolving towards CV syllables, as the reviewer might believe, should we say that all CC onsets in English are outliers? Clearly we do not, because CC onsets are part of English grammar here and now.

4. Explaining the spotty-data problem

Why are so many possible words not used in English? One might suspect that there are phonological constraints that rule out most of the disyllabic words, but this is unlikely: there are no known phonological constraints that would rule out 99% of the disyllabic combinations. Instead, the answer in my view is that a language simply does not need many morphemes, which make up words (Duanmu 2008).

Let us consider the size of morpheme inventories in English and Chinese. For simplicity and consistency in calculation, I take the number of morphemes in Chinese to be roughly the same as the number of characters. This method ignores homographs. For example, the character 白 can mean 'white' or 'in vain' but it is counted as one morpheme. The analysis also over-counts disyllabic morphemes, although there are not many in Chinese. For example, 蜻蜓 'dragonfly' and 玛瑙 'amber' are each one morpheme, but they are counted as two morphemes each, even though the parts have no meaning by themselves.

For English, I take the number of morphemes to be roughly the same as the number of words that are labeled as single morphemes in CELEX (excluding proper names). This method will count homographs because CELEX lists them separately. For example, *bank* (of money) and *bank* (of river) are listed separately and will be counted as two morphemes. However, the analysis excludes bound morphemes, such as *bio-*, *pre-*, *-ology*, *-er* and *-ly*. The undercount of bound morphemes in English is compensated by the inclusion of homographs, which are excluded in Chinese. Therefore, the overall effects of the counting method probably balance out for the two languages.

In both languages, zero derivations (i.e. a change of word category without an overt affix) are excluded. For example, in English *dry* (adjective) is included but *dry* (verb) is not. Similarly, in Chinese 干 is counted once, although it can be a verb 'to dry', an adjective 'dry', or a noun 'dried food'.

I use two electronic corpora for the comparison, Da (2004) for Chinese and CELEX for English. The basic information of the corpora is given in (24).

(24)		Chinese	English
	Corpus	Da (2004)	CELEX (Baayen et al 1993)
	Size	259 million characters	18 million words
	Morphemes	12,041 character types	7,401 monomorphemic words

The English corpus has fewer morphemes because it covers modern English only, while the Chinese corpus covers both classic and modern texts. In addition, many characters in the Chinese corpus are rarely used. If we ignore uncommon morphemes, the similarity between the languages becomes more evident. To see it, let us consider the coverage of character or word tokens in each corpus. The data are shown in (25), up to the 7,000th most frequent morpheme. The Chinese calculation is made by Da (2004). The English calculation is made by me.

(25) Cumulative corpus coverage by the number of most frequent morphemes

Most frequent	Chinese coverage	English coverage
1,000	86.1740%	87.3571% (*wise*, 723)
2,000	95.5529%	94.2505% (*liquid*, 204)
3,000	98.3248%	97.2358% (*leap*, 78)
4,000	99.3046%	98.6762% (*loom*, 35)
5,000	99.7321%	99.4708% (*tankard*, 16)
6,000	99.9268%	99.8682% (*clunk*, 5)
7,000	99.9802%	100.0000% (*gull* (verb), 0)

The first 1,000 most frequent characters in Chinese cover 86% of all character tokens and the first 1,000 most frequent English morphemes, which ends at *wise* (occurring 723 times), cover 87% of all word tokens. In both languages, the first 4,000 most frequent morphemes cover 99% of all occurring tokens, and the first 6,000 most frequent morphemes cover 99.9% of all occurring tokens. The bottom 454 morphemes in English, such as *asp* (noun), *barm* (noun), and *gull* (verb), do not occur in the frequency corpus (the frequency corpus was one of the sources from which the CELEX lexicon was gathered), but it is reasonable to assume that they are infrequent and do not affect the overall results. In any case, in both Chinese and English, morphemes beyond the first 6,000 most frequent ones cover just 0.1% of all occurrences.

It is unclear how many morphemes are used in other languages. However, it is reasonable to assume that they are unlikely to be much larger than those in English and Chinese, because English is used world wide and has borrowed many words from other languages, and the Chinese corpus was based on not only modern usage but also a large amount of texts from classic literature. If so, in most languages the number of morphemes needed is just a very small fraction of possible words available.

So if a language only needs 1% (or a few percent) of all possible words, which ones would be chosen? There are two possibilities: either the words are chosen more or less arbitrarily, or they are chosen according to phonological principles. Our examination of syllables in English and Chinese shows that word forms can often be chosen arbitrarily. If so, it can sometimes be difficult to make phonological generalizations in a language, because we cannot be sure whether the generalizations are real or merely artifacts.

5. What can phonologists do?

Despite the spotty-data problem, there are plenty of things phonologists can do. First of all, there are often clear rules or generalizations that should be included in the description of a language. For example, American English has a rule to flap [t, d] between vowels, Standard Chinese has a rule to change Tone 3 to Tone 2 when the following tone is Tone 3, and Shanghai Chinese has a rule to delete the underlying tone from the second word of a disyllabic compound. These are features that distinguish one language from another.

The phonology of a language should also describe which sounds and words are used in the language and, based on such inventories, what types of sound sequences are used in syllables. In addition, the phonology should also describe which sounds and words are used frequently and which occasionally, and based on such information one can predict which non-words would sound more acceptable to native speakers than other non-words, although it is another question whether more acceptable non-words will actually be adopted earlier than others.

Phonologists can also look for universals that hold for all languages. Ironically, this may sometimes be easier than looking for generalizations of a particular language. The reason is that, because of the spotty-data problem, there are often not enough data in a given language, yet if we look at all languages, we have a lot more data for making generalizations.

6. Summary

A language often uses only a few thousand morphemes, which can often be distinguished by just a small fraction of all possible combinations of consonants and vowels in its sound inventory. Therefore, many possible syllables or word forms will not occur, not because they violate any phonological constraint, but because the language simply does not need so many syllables or word forms. The paucity of occurring forms (i.e. the spotty-data problem) makes it hard sometimes to figure out a phonological rule or constraint. Still, there are often clear language-particular generalizations to make, as well as universal generalizations that are true for all languages.

REFERENCES

Abercrombie, David. 1967. Elements of general phonetics. Chicago: Aldine.

Baayen, R. Harald, Richard Piepenbrock, and L. Gulikers. 1993. The CELEX lexical database (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Blevins, Juliette. 2003. Evolutionary phonology: the emergence of sound patterns. Cambridge: Cambridge University Press.

Burzio, Luigi. 1994. Principles of English stress. Cambridge, UK: Cambridge University Press.

Cheng, Chin-Chuan. 1973. A synchronic phonology of Mandarin Chinese. Monographs on linguistic analysis no. 4. The Hague: Mouton.

Chomsky, Noam. 1957. Syntactic Structures. The Hague: Mouton.

Chomsky, Noam. 1981. Lectures on government and binding. Dordrecht: Foris.

Chomsky, Noam, and Morris Halle. 1968. The sound pattern of English. N.Y.: Harper and Row.

Clements, G. N., and Samuel Jay Keyser. 1983. CV phonology: a generative theory of the syllable. Cambridge, Mass.: MIT Press.

Da, Jun. 2004. Chinese text computing. Murfreesboro: Department of Foreign Languages and Literatures, Middle Tennessee State University. <http://lingua.mtsu.edu/chinese-computing/>

Davis, Stuart. 1988. Topics in syllable geometry. New York: Garland.

Duanmu, San. 2000. The phonology of Standard Chinese. Oxford: Oxford University Press.

Duanmu, San. 2008. Syllable structure: How different can it be in human languages? Oxford: Oxford University Press. (Expected)

Frisch, Stefan A., N. R, Large, and David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. Journal of Memory and Language 42.4: 481-496.

Giegerich, Heinz. 1992. English phonology. Cambridge, UK: Cambridge University Press.

Gimson, A. C. 1970. An introduction to the pronunciation of English. 2nd edition. New York: St. Martin’s Press.

Halle, Morris. 1962. Phonology in generative grammar. Word 18:54-72.

Halle, Morris, and Jean-Roger Vergnaud. 1987. An essay on stress. Cambridge, Mass.: MIT Press.

Hammond, Michael. 1999. The phonology of English: a prosodic Optimality Theoretic approach. Oxford: Oxford University Press.

Jones, Daniel. 1950. The pronunciation of English. 3rd edition. Cambridge, England: Cambridge University Press.

Kahn, Daniel. 1976. Syllable-based generalizations in English phonology. Doctoral dissertation, MIT, Cambridge, Mass.

Ladefoged, Peter. 2001. Vowels and consonants: an introduction to the sounds of languages. Malden, MA: Blackwell.

Lin, Yen-hwei. 1989. Autosegmental treatment of segmental processes in Chinese phonology. Doctoral dissertation, University of Texas, Austin.

Myers, James, and Jane Tsay. 2005. The processing of phonological acceptability judgments. Proceedings of Symposium on 90-92 National Science Council Projects, pp. 26-45. Taipei, Taiwan.

Pierrehumbert, Janet. 1994. Syllable structure and word structure: a study of triconsonantal clusters in English. In Phonological structure and phonetic form, Papers in Laboratory Phonology III, ed. Patricia A. Keating, 168-188. Cambridge, England and New York, NY: Cambridge University Press.

Prince, Alan, and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Ms., Rutgers University and University of Colorado.

Pulgram, Ernst. 1970. Syllable, word, nexus, cursus. [Janua linguarum Series minor. 81.] The Hague: Mouton.

Selkirk, Elisabeth. 1982. The syllable. In The structure of phonological representations (Part II), ed. Harry van der Hulst and Norval Smith, 337-83. Linguistic Models 2. Dordrecht: Foris. Abbreviated and reprinted in John Goldsmith 1999, 328-50.

Steriade, Donca. 1999. Alternatives to syllable-based accounts of consonantal phonotactics. In Osamu Fujimura, Brian D Joseph, and Bohumil Palek, eds., Proceedings of LP'98: Item Order in Language and Speech (Columbus, the Ohio State University, September 15-20, 1998), Vol. I, 205-245. Prague: Karolinum Press (Charles University in Prague).

Treiman, Rebecca, and Catalina Danis. 1988. Syllabification of intervocalic consonants. Journal of Memory and Language 27.1: 87-104.

Wells, John Christopher. 1990. Syllabification and allophony. In Studies in the pronunciation of English, A commemorative volume in honour of A.C. Gimson, ed. Susan Ramsaran, 76-86. London and New York: Routledge.

Yip, Moira. 1988. The obligatory contour principle and phonological rules: a loss of identity. Linguistic Inquiry 19.1: 65–100.

Zhang, Xinting. 2007. Lexical Decision in Standard Chinese. Ms., Department of Linguistics, University of Michigan.