Interfaces in Chinese Phonology, 261-278
2008-8-008-011-000098-1

The "Spotty-Data Problem" and Boundaries of Grammar*

San Duanmu
University of Michigan

The main job of a linguist is to study grammar—either universal grammar, which includes common properties shared by all languages, or the grammar of a particular language. The job would be easier if the boundaries of grammar were clear, but often they are not. A few examples are taken from phonology, with a focus on what I call the "spotty-data" problem, which refers to the fact that we often do not have enough data to determine where the boundaries of grammar are.

Key words: spotty-data problem, grammaticality, gradient judgment, phonotactic frequency, potential words

1. Introduction: generative grammar

Chomsky (1957) proposes 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.

* Matthew Chen has shown many people, including myself, how to do phonology. His works, always clear, theoretically pertinent, and descriptively grounded on solid data, have set a high standard for others to follow. I am pleased to contribute to this volume in his honor, in order to continue the quest for a better understanding of phonology, grammar, and the human mind. This paper benefited from the comments of two anonymous reviewers and the editors of this volume, to whom I am grateful.

(2) N → A-*ness*
A → N-*less*
Grammatical words, which can be infinitely long:
red, 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. Later revisions have replaced rules with principles and parameters (Chomsky 1981), but the essence of the proposal remains the same: the grammar is well defined; so are the structures it generates.

2. Grammar and intuition

Chomsky (1957) also proposes that, because the speaker of a language has an implicit 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. Similarly, Halle (1962) argues that speaker intuition can be used to judge whether a sound sequence is well formed, even if it is not a real word. For example, [bɪk], [nɪs], and [bnɪk] are (or were) not real English words, but the first two are possible words while the third is not.

However, recent studies (e.g. Frisch et al. 2000, Myers & Tsay 2005, and Zhang 2007) have shown that speaker judgment on possible words is not always clear-cut. Similarly, many studies have noted that consistent judgment on syllable boundaries can be hard to obtain (e.g. Gimson 1970, Treiman & Danis1988, Giegerich 1992, Hammond 1999, Steriade 1999, Blevins 2004). Linguists do not always agree on syllable judgment either. An example is shown in (3), where a dot indicates a syllable boundary and [t̲] is a single [t] that belongs to both the first and the second syllable.

(3) Syllable boundary in *city*

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

Giegerich (1992) proposes that speaker judgment can be probed with a **syllable-repetition** test, which may resolve the ambiguity in (3). In the test, speakers are asked

to pronounce each syllable of a word twice. One can begin with words whose syllable boundaries are unambiguous, such as *after*, which is pronounced as *af-af-ter-ter*. More difficult words can then be presented to see how speakers deal with them. One such word is *apple*, whose output is reported to be *ap-ap-ple-ple*. Giegerich argues that the result means that *apple* is syllabified as *ap-ple*, where the [p] is ambisyllabic, as proposed by Pulgram (1970) and Kahn (1976). Similarly, the syllabification of *city* should be *cit̲y*, too.

There are several questions about the syllable-repetition test. First, if the output for *apple* is *ap-ap-ple-ple*, instead of *ap-ap-le-le*, does it mean that the syllables are *ap-ple*, or does it mean that *le-le* [l̩-l̩] is an unusual sequence of syllables which the speaker is trying to avoid? Second, consider the word *text*. If the result is *tek-tek-st-st*, one might conclude that the syllable is [tɛk] and [st] is outside of it. However, if the result is *text-text*, does it mean that the syllable is [tɛkst], or does it mean that the speaker is trying to avoid repeating a non-syllablic cluster [st]? Finally, Giegerich does not discuss whether speaker judgment is always clear. My own test with some native speakers shows that the judgment can vary. For example, the output for *city* can be *cit-cit-ty-ty* or *ci-ci-ty-ty*. Therefore, the test does not seem to provide conclusive answers.

But suppose there is a total lack of speaker intuition for syllable boundaries, should we conclude that there are no syllable boundaries? The answer in my view is no. The reason is 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. The lack of intuitive judgment on linguistic structures simply means that linguists have to work harder in figuring out patterns of grammar.

3. Grammar and phonetic data

The phonological rules of a grammar often depend on the phonetic transcription of the data, or the interpretation of the transcription. Similarly, assumptions about phonological rules can affect one's transcription of the data. As an example, consider allophonic variations of the mid vowel in Standard Chinese, shown in (4) and (5). The analysis is similar to what is offered by Cheng (1973), which was also adopted, with minor variations, by Chen (1984) and Steriade (1988).

(4) Transcription of the variation of the mid vowel in Standard Chinese
[e]: [ei], [je], [ɥe] (*[ɥö]), [wei] (*[woi])
[o]: [ou], [wo], [jou] (*[jeu])
[ə]: Elsewhere, e.g. [kə]

(5) Analysis of (4)
  a. Underlying form: [ə] (unspecified for [back])
  b. Rules:
    1. a. [ə] → [-back] / __[-back]
      ([ə] becomes [e] before [i])
    b. [ə] → [+round] / __[+round]
      ([ə] becomes [o] before [u])
    2. [ə] → [-back] / [-back]__
      ([ə] becomes [e] after [i] or [ɥ])
    3. [ə] → [+round] / [+round]__
      ([ə] becomes [o] after [w])
  c. Rule ordering: Rule 1 before Rule 2; Rule 2 before Rule 3

In this analysis, Standard Chinese has one mid vowel, which is unspecified for the feature [back] and has three variations, predictable from the environments. To account for the variations, we can propose three ordered rules for the grammar.

However, if we transcribe the data slightly differently, we end up with different rules, and hence a different grammar. The alternative is shown in (6) and (7), proposed by Wang (1993).

(6) Transcription of the variation of the mid vowel in Standard Chinese
  [e]: [jeː], [ɥeː]
  [o]: [woː]
  [ə]: Elsewhere, e.g. [kə], [əi], [wəi], [əu], [jəu]
(7) Analysis of (6)
  a. Underlying form: [ə]
  b. Rules:
    1. [əː] → [-back] / [-back]__
      ([əː] becomes [eː] after [i] or [ɥ])
    2. [əː] → [+round] / [+round]__
      ([əː] becomes [oː] after [w])
  c. Rule ordering: Rule 1 before Rule 2

In the new analysis, the mid vowel remains [ə] when it is short, which is the case in [əi], [wəi], [əu], and [jəu], where the rhyme is a diphthong. The mid vowel changes only when it is long, which happens in a stressed open syllable. As a result, we have just two rules, and they apply to long [əː] only.

The example shows that it is not always easy to determine how many rules, or what

kind of rules, there are in a language, because the differences in phonetic transcriptions are rather subtle. In addition, whether we use rule-based analyses or constraint-based analyses of the Chinese data (Lin 1997), the point remains the same.

A similar problem exists with the transcription and analysis of diphthongs in English. Consider two approaches, shown in (8) and (9).

(8) Transcription of diphthongs in English: [eɪ, oʊ, aɪ, aʊ]
Constraint: Diphthongs end in a lax high vowel

(9) Transcription of diphthongs in English: [ei, ou, ai, au]
Constraint: Diphthongs end in a tense high vowel

The difference between (8) and (9) is again rather subtle. Supporters of (8) may argue that, phonetically, the end point of a diphthong does not quite reach that of a tense high vowel, but often ends at that of a lax high vowel. On the other hand, supporters of (9) may argue that, phonologically, the intended end target of a diphthong could be a tense high vowel, although owing to lack of time, the target is not always reached. A third approach is also possible, which employs the notion of underspecification (Steriade 1987, Archangeli 1988). Because there is no contrast in the feature [tense] in the second half of a diphthong, the high vowel is unspecified for [tense]. The analysis is shown in (10), where [I] indicates a high front vowel unspecified for [tense] and [U] indicates a high back round vowel unspecified for [tense].

(10) Observation: no contrast in [ei]-[eɪ], [ou]-[oʊ], [ai]-[aɪ], or [au]-[aʊ]
Transcription: [eI, oU, aI, aU]
Constraint: [tense] is not used in the second half of a diphthong

Indeed, we can go even further. We observe that there is no contrast in [eI]-[ɛI] or [oU]-[ɔU], and so we can conclude that there is no contrast in [tense] in the first part of a diphthong either. Therefore, we can transcribe the diphthongs are [EI] and [OU], and propose a constraint that [tense] is not used at all in diphthongs (Duanmu 2008).

In the analyses of the Chinese mid vowel the choice does not depend so much on the accuracy of phonetic data, but on the interpretation of the data. Similarly, in the analyses of English diphthongs, or diphthongs in any language, the choice does not depend on the accuracy of phonetic data either, but again on the interpretation of the data. Such examples show that simple measurements or experiments cannot always resolve the question of where the boundaries of a grammar lie.

4. The "spotty-data" problem

Let us consider another problem, which I call the **spotty-data** problem (Duanmu 2008). It refers to the fact that often there are not enough data for making reliable generalizations, even if we examine the entire lexicon of a language. To see the problem, 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 (11), where V is a short (lax) vowel.

(11) 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, tʃ, dʒ, 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, tʃ, dʒ, m, n, ŋ, l, r]
Total CVC: 2,415

American English has 24 consonants, including the affricates [tr, dr, tʃ, dʒ]. 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 (12), based on the CELEX lexicon (Baayen et al. 1993).

(12) 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),

[1] Whether [tr, dr] are single sounds (affricates) or clusters of two each does not affect the point being made here. However, a reviewer requests explanations for treating [tr, dr] as affricates. Phonetically, [tr, dr] are similar to affricates, in the sense that there is clear frication after the release of the stop, a view shared by many phoneticians, such as Jones (1950), Abercrombie (1967), Gimson (1970), and Wells (1990). Phonologically, two questions may be raised. First, an affricate is usually made of a stop and a fricative, yet [r] is an approximant. How then can [t] + [r] or [d] + [r] create an affricate? The answer is that there is already evidence that stop + approximant can make an affricate, such as [t] + [j] → [tʃ] in *get you* and [d] + [j] → [dʒ] in *did you*. The second question is, 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.

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 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.

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. Rather, the real 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 都 can mean 'capital' or 'all' but it is counted as one morpheme.[2] The analysis also over-counts disyllabic morphemes, although there are not many in Chinese. For example, 蜻蜓 'dragonfly' and 瑪瑙 '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 乾 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 (13).

[2] As a reviewer points out, in simplified characters, there are more homographs. For example, 'dry' and 'to do' are represented with the same graph 干 in simplified characters but distinct graphs 乾 and 幹 in traditional characters.

(13)

	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 (14), up to the 7,000th most frequent morpheme. The Chinese calculation is made by Da (2004). The English calculation is made by me.

(14) 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 or 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. It is often possible to look at the lexicon of a language and make various phonological generalizations. However, without knowing whether a lexicon is an arbitrary or systematic collection of words, we cannot be sure whether the generalizations are real or merely artifacts.

5. Holes and outliers in syllable patterns

If a language only needs a fraction of available syllables, it has considerable freedom to choose which ones to use. The selected syllables may be subject to some markedness constraints—constraints that favor certain sound combinations over others. For examples, languages can choose which consonants are used in the coda of a syllable, as seen in (15).

(15) Coda consonant in English, Cantonese, and Standard Chinese (SC), where S = fricative, TS = affricate, T = stop, and N = nasal

	[l]	TS	S	T	N
English	+	+	+	+	+
Cantonese	–	–	–	+	+
SC	–	–	–	–	+

Similarly, a Cantonese syllable generally disallows two labial sounds (Yip 1988); the only coda consonants in Standard Chinese are [n] and [ŋ]; and syllables in Shanghai generally do not have a coda (Duanmu 1999). Such patterns have led to the view that the phonological structures of a language can be precisely described by a set of rules, as proposed by Halle (1962). A similar view is held by Prince & Smolensky (1993).

On the other hand, because a language does not need many syllables, there are often **holes** in the distribution—syllables that could have been used but are not. For example, Standard Chinese uses [məu] 'strategy', [fəu] 'not', and [p^həu] 'dissect', but no [pəu], which seems to be a hole. Similarly, English has at least 59 **productive** onsets, shown in (16)-(19). 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.

(16) 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)

(17) C onsets
Productive (22): p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, h, tr, dr, tʃ, dʒ, m, n, l, r
Unproductive (1): ʒ

(18) CC onsets
Productive (30): 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 (36): 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*)

(19) CCC onsets
Productive (6): str (*string*), skr (*screen*), skw (*square*), spr (*spring*), spl (*splash*), skl (*sclerosis*)
Unproductive (4): 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 (20).

(20) 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 the string C+ [lɪl] (Clements & 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 holes.

Besides holes, there are often **outliers**—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* [skeə˞][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].[3]

If we always know what are holes and what are outliers, 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 (a hole) or simply ungrammatical. Our decisions in such cases would lead to fairly different versions of English grammar. For example, (21) and (22) show two ways to treat *Tswana* and *scherzo*.

(21) Decision: *Tswana* and *scherzo* are outliers (not good words) in English.[4]
Generalizations: English words and syllables cannot start with [ts].
[tsæt], [tsɪl], … are impossible (ungrammatical) words in English.

(22) 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 (holes) in English.

Similarly, the decision on words like *sphere* can lead to two analyses, shown in (23) and (24).

(23) 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.

[3] 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?

[4] 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 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.

(24) 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 (holes) 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 holes or potential words in English.

As another example, consider occurring and non-occurring monosyllables with VC rhymes in English, shown in (25).

(25) 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 (26). The number of occurring onsets for a rhyme is in parentheses and the CELEX [O] can be [ɒ] or [ɑ] in American English.

(26) 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), [ɪtʃ] (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), [ætʃ] (10), [ɛdʒ] (10), [ɪd] (10), [On] (10), [Oŋ] (10), [ʌs] (10), [ʌdʒ] (9), [ʌŋ] (9), [ɛg] (8), [ɛtʃ] (8), [ɪs] (8), [Of] (8), [Otʃ] (8), [ʊk] (8), [ʌl] (8), [æs] (7), [ɛf] (7), [ɛm] (7), [Ol] (7), [Om] (7), [ʌtʃ] (7), [ɛp] (6), [ɪʃ] (6), [Oθ] (6), [ʊd] (6), [Odʒ] (5), [ʌp] (5), [ʌv] (5), [æl] (4), [ɛb] (4), [ɛʃ] (4), [ɪdʒ] (4), [ɪθ] (4), [ɪv] (4), [ʊl] (4), [ʊʃ] (4), [ʊt] (3), [ʌz] (3), [ædʒ] (2), [æf] (2), [æv] (2), [æz] (2), [ɛθ] (2), [ʊtʃ] (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 holes, although even the most frequent rhymes hardly occur with half of the onsets. The hard 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? The answer again leads to two analyses, shown in (27) and (28).

(27) Decision: *with* is an outlier (not good word) in English.
Generalizations: [ɪð] is not a possible rhyme in English.
[mɪð, nɪð, tɪð,…] are impossible (ungrammatical) words in English.

(28) 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 (holes) in English.

If *with* is an outlier (phonologically bad) 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 holes 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 happens not to have used any (or many) such words, but is in principle open to their use? The two options are shown in (29) and (30).

(29) Decision: [pʌm] 'pump' is an outlier (not good word) in Cantonese.
Generalizations: Cantonese does not allow any syllable that has a labial onset and a labial coda.
[pau, mau, …] are impossible (ungrammatical) words in Cantonese.

(30) 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 (holes) 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, which generally lacks syllables that have a 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 real 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, 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 holes?

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 **holes** 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.

The judgment of native speakers might not be much help to solve the problem of holes and outliers, because their judgment might simply reflect whether a word is or is not in their language, or how similar it is to an existing word, or how many existing words it is similar to, rather than what phonological principles are. Clearly, the issue of **holes** and **outliers** adds additional difficulty to the determination of boundaries of grammar.

6. Universal grammar and particular grammars

Although the boundaries of grammar are often hard to determine, it does not mean that we cannot work on the grammar of a language. There are, very often, clear rules or generalizations that must 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.

A grammar should also describe which sounds and words are used in the language

and, based on such inventories, what types of syllables are used. In addition, a grammar 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 used earlier than others.

In many ways the study of a language is like the study of the course of a river. At any point of time the course of a river is fairly well defined intuitively (e.g. those parts covered by flowing water), so is a language (e.g. sounds, words, and expressions that are or can be used). However, over time the notion of a river becomes less well defined, so is that of language. For example, should a past or occasional course be seen as part of a current river? Should a past or occasional usage be seen as part of a current language? Should a possible future course (in case there is a huge flood) be seen as part of a current river? Should a possible future word be seen as part of a current language? Answers to such questions may affect certain generalizations about a river or a language. For example, some rivers may never be wider than a mile, unless you include occasional floods. Similarly, Standard Chinese does not have a word [si], but many speakers can now pronounce ABC as [ei, bi, si] (instead of [ɕi] or [sei] for C), or CCTV (for 'China Central TV') as [si, si, ti, vi] (instead of [ɕi, ɕi, ti, wei] or [sei, sei, ti, wei]). If we include ABC and CCTV in the vocabulary, then Standard Chinese has the syllable [si] (and [vi]), but otherwise it does not.

Some generalizations are conditional on a particular river or language. For example, whether a course flows eastward or westward, whether a course is wider than a mile, or whether the water is clear, brown, or green, depends on what river it is. Similarly, whether a language uses the syllable [si], or the sound [θ], depends on what language it is. The course of a river is the result of many accidental events; so is a language. In addition, the future course of a river is predictable only if you know all the factors involved, including weather and human actions. Similarly, the future of a language is predictable only if you know all the factors that affect language change, including social and cultural interactions.

On the other hand, there are generalizations that are unconditional and true for all rivers, regardless of the environment or time. For example, all rivers flow from a higher place to a lower place. Similarly, there are generalizations that are true for all languages, regardless of time. For example, all natural languages use consonants and vowels to make words, all consonants and vowels can be represented by a dozen or so distinctive features, and (most linguists believe that) all natural languages have the word categories noun, verb, and adjective. Such general principles should be an indispensable part in the study of rivers (geological principles) or languages (universal grammar), even though the principles may ultimately blend away into other and more fundamental disciplines of science.

7. Summary

A language often only uses a few thousand morphemes and, if there are disyllabic morphemes or a certain amount of homophones, there will be a much smaller number of syllables. Therefore, it is often hard to tell whether an unused syllable is simply not needed (a potential word) or ruled out for phonological reasons (an impossible word or an ungrammatical form). Still, while the boundaries of a grammar are not always clear, 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.

Archangeli, Diana. 1988. Aspects of underspecification theory. *Phonology* 5:183-207.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. 1993. The CELEX lexical database. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Blevins, Juliette. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge & New York: Cambridge University Press.

Burzio, Luigi. 1994. *Principles of English Stress*. Cambridge & New York: Cambridge University Press.

Chen, Matthew Y. 1984. Abstract symmetry in Chinese verse. *Linguistic Inquiry* 15.1: 167-170.

Cheng, Chin-Chuan. 1973. *A Synchronic Phonology of Mandarin Chinese*. The Hague: Mouton.

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

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Clements, George N., and Samuel Jay Keyser. 1983. *CV Phonology: A Generative Theory of the Syllable*. Cambridge: 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. 1999. Metrical structure and tone: evidence from Mandarin and Shanghai. *Journal of East Asian Linguistics* 8.1:1-38.

Duanmu, San. 2000. *The Phonology of Standard Chinese*. Oxford & New York: Oxford University Press.

Duanmu, San. 2008. *Syllable Structure: How Different Can It Be in Human Languages?* Oxford & New York: Oxford University Press. (Expected)

Frisch, Stefan A., Nathan 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 J. 1992. *English Phonology*. Cambridge & New York: Cambridge University Press.

Gimson, Alfred 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* 18.1-2:54-72.

Halle, Morris, and Jean-Roger Vergnaud. 1987. *An Essay on Stress*. Cambridge: MIT Press.

Hammond, Michael. 1999. *The Phonology of English: A Prosodic Optimality Theoretic Approach*. Oxford & New York: Oxford University Press.

Jones, Daniel. 1950. *The Pronunciation of English* (3rd edition). Cambridge: Cambridge University Press.

Kahn, Daniel. 1976. *Syllable-based Generalizations in English Phonology*. Cambridge: MIT dissertation.

Lin, Yen-Hwei. 1989. *Autosegmental Treatment of Segmental Processes in Chinese Phonology*. Austin: University of Texas dissertation.

Lin, Yen-Hwei. 1997. Assimilation in Mandarin and feature class theory. Paper presented at the 9th North American Conference on Chinese Linguistics (NACCL-9). Victoria, Canada.

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

Pierrehumbert, Janet. 1994. Syllable structure and word structure: a study of triconsonantal clusters in English. *Phonological Structure and Phonetic Form*, ed. by Patricia A. Keating, 168-188. Papers in Laboratory Phonology 3. Cambridge & New York: Cambridge University Press.

Prince, Alan, and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Manuscript. New Brunswick: Rutgers University; Boulder: University of Colorado.

Pulgram, Ernst. 1970. *Syllable, Word, Nexus, Cursus*. Janua linguarum Series minor, 81. The Hague: Mouton.

Selkirk, Elisabeth. 1982. The syllable. *The Structure of Phonological Representations* (Part II), ed. by Harry van der Hulst & Norval Smith, 337-83. Linguistic Models 2. Dordrecht: Foris.

Steriade, Donca. 1987. Redundant values. *CLS* 23.2:339-362. Chicago: Chicago Linguistic Society.

Steriade, Donca. 1988. Review of CV Phonology, by George N. Clements & Samuel Jay Keyser. *Language* 64.1:118-129.

Steriade, Donca. 1999. Alternatives to syllable-based accounts of consonantal phonotactics. *Proceedings of LP'98: Item Order in Language and Speech* (Columbus, the Ohio State University, September 15-20, 1998), Vol. 1, ed. by Osamu Fujimura, Brian D. Joseph & Bohumil Palek, 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.

Wang, Jenny Zhijie. 1993. *The Geometry of Segmental Features in Beijing Mandarin*. Newark: University of Delaware dissertation.

Wells, John Christopher. 1990. Syllabification and allophony. *Studies in the Pronunciation of English: A Commemorative Volume in Honour of A. C. Gimson*, ed. by Susan Ramsaran, 76-86. London & 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*. Manuscript. Ann Arbor: University of Michigan.

Department of Linguistics
University of Michigan
440 Lorch Hall
611 Tappan Street
Ann Arbor, MI 48109-1220
USA
duanmu@umich.edu