Phonetic Variation of Coronals in English Infant-Directed Speech:

A Large-Scale Corpus Analysis

Ekaterina A. Khlystova

Abstract

Phonetic variation poses a challenge for language learners tasked with identifying the abstract sound categories (phonemes) and positional allophony of their target language(s). Yet we know relatively little about the actual degree of phonetic variability in IDS and how this variation is structured. In this study, we set out to provide a more holistic understanding of what infants hear by quantifying the extent of variability in the pronunciation of some of the most frequent sound categories of English: coronals (/t/, /d/, /s/, /z/, and /n/). We further examine the degree to which this variation is expected based on English phonotactics. We sampled IDS from the longitudinal Providence Corpus (Demuth et al. 2006) which contains recordings of 5 typically-developing, monolingual, English-speaking 1- to 3-year-olds interacting with their caregivers at home during everyday activities. These utterances were force-aligned (Rosenfelder et al. 2014) according to orthographic transcripts to generate segmental boundaries automatically. We then checked and phonetically annotated ~7,000 utterances containing 31,245 coronal segments.

We found that overall, canonical variants of /t/ are in the minority (39%) whereas /s/ is overwhelmingly canonical (98%); further, almost every segment had more canonical instances in word-initial compared to word-final position. We also examined the distribution of expected variants based on English phonotactics against the observed variants for /t/ and /d/, two segments that are the most variable. While most variants had high counts of matching observed and expected variants, we also find that unexpected variants are common. The results of the current study help provide an understanding of the full extent of variation in naturalistic IDS. We discuss the implications of these results for theoretical and computational models of morphological and phonological acquisition.

1 | Introduction

1.1 General Introduction

Typically developing children learn their native language(s) at a spectacular rate – most children are highly competent users of their language by the time they are 5-6 years old. The magnitude of this tremendous feat is further highlighted by the fact that the linguistic input children hear is variable, contains overlapping sound categories, and frequently demonstrates semantic or syntactic ambiguity. One of the first tasks with which a child is faced is learning the sound systems of their language(s). This is a vital task, as these systems serve as the foundational building blocks for later acquisition of morphology, syntax, semantics, and pragmatics. Furthermore, the acquisition of speech sound categories entails learning not only their canonical, or prototypical, instances but also the less typical ones. This variability poses a significant challenge to the young learner as they must simultaneously learn the prototypical form and other possible variants. Further, because this variation is language-specific, all of this must be learned solely from their linguistic input: infant directed speech (IDS).

We know from research on adult-directed speech (ADS) that some of the variation within categories is highly systematic and arises as a result of applying phonological processes. One example of such variation is positional allophony, such as intervocalic tapping ("butter" [bʌɾə˞]), syllable-initial aspiration of voiceless stops ("pit" [pʰɪt]), among many others. On the other hand, some modifications are not as systematic and are the result of high rates of speech, such as syllable or segment deletion (Johnson 2004), vowel reduction/deletion, coda deletion (Bell et al. 2003), schwa deletion (Dalby 1986, Patterson et al. 2003) and voicing assimilation (Ernestus et al. 2006, Snoeren et al. 2006). While this sort of variation is extensively studied in adult speech, relatively little is known about the extent of phonetic variation in IDS.

1.2 The phonological learning problem.

One of the biggest challenges in phonological acquisition is extracting the phonemic form from a large set of surface variants, a seemingly impossible task – and yet infants have detailed knowledge of the phonemes of their language(s) by the time they are 12 months old (Werker and Tees 1984, Polka and Werker 1994, Kuhl et al. 2008 among others). In one proposal, the phonological learning problem is simplified because of the properties of IDS; the input infants receive is highly canonical, and is consequently less "noisy" due to variants. For this reason, early research on phonetic variation in IDS aimed primarily at comparing the degree of phonetic variability in infant input to that of ADS to explain the precocious success infants exhibit in extracting phonemes from their language(s) input.

1.2.1 Notes on terminology.

Before discussing the previous literature on variation, it is necessary to clarify the terminology used in past work and the terminology to be used in this paper. In research about speech directed to infants and adults, the term "canonical" is generally equated with "phonemic". That is to say, learning the canonical form for a segment is considered equivalent to learning one specific abstract phoneme from a set of phonetic variants. In this study, we adopt the same convention, and will use the terms canonical and phonemic interchangeably unless explicitly stated otherwise. Similarly, we will adopt the term "faithful" as referring to those variants that surface as the underlying phonemic form – for example, a /t/ surfacing as [t]. Lastly, we adopt the convention that "allophone" refers to all other variants of an underlying phoneme.

1.2.2 Previous work on phonetic variation in IDS.

In early descriptions, IDS (also known as "baby talk" or "motherese") was characterized as having more canonical segments compared to ADS (Ferguson 1964, Bernstein Ratner 1984). "Simplified" consonant and vowel categories in IDS were thought to facilitate language acquisition and hold the attention of infants (Ferguson 1964). Consistent with this idea, early studies on IDS reported that vowel formants were more widely dispersed, decreasing the overlap between vowel categories (Bernstein Ratner 1984; Burnham et al. 2002; Kuhl et al. 1997). More recently, Dilley, et al. (2014) examined regressive place assimilation in English in recordings of parents reading to their children and found that IDS contained more canonical forms than ADS. However, it is important to note that these forms were read, and read speech is typically more careful and less variable than spontaneous speech. This characterization of IDS as clear speech is consistent with the idea of variation as noise, such that hearing IDS allows infants to postpone the need to learn the distribution of phonetic variants until after the phonemic categories are fully extracted and acquired. But the "enhancements" typically cited as beneficial modifications in IDS – such as increased formant distance between vowels (Kuhl et al. 1997) or fewer assimilated consonants (Dilley et al. 2014) – tend to be unreliable, and even when present, may not be beneficial to learning (Cristia and Seidl 2014, Martin et al. 2015).

While IDS as a simplified, pedagogical manner of speech is a compelling idea, recent work has suggested that IDS may instead exhibit a degree of variation on par with that observed in ADS. Lahey and Ernestus (2014) examined a corpus of naturalistic speech and found that IDS contains as much reduced speech as ADS. However, their study only compared pronunciation variation in two highly frequent lexical items, calling into question the generalizability of these results. Buckler, Goy, and Johnson (2018) examined place assimilation in speech to English-learning 18-

month-olds and found that “infant directed speech contains as many non-canonical realizations of words in place assimilation contexts as adult-directed speech” (p. 45). In fact, there are even some early reports of IDS having more phonological reduction than ADS, posing an additional challenge to the learner (Shockey and Bond 1980). Further, a large-scale corpus analysis of spontaneous Japanese speech showed that IDS had a small but significant tendency to have less clear contrasts than in ADS (Martin et al. 2015).

Phoneme acquisition is not the only aspect of language acquisition that must occur – learning word forms also relies on generating an abstract phonemic and acoustic entry for each item. In a corpus study of Japanese speech, Guevara-Rukoz et al. (2018) investigated whether IDS can facilitate word-form learning when compared to ADS by examining words at both an acoustic and phonological level. Specifically, they used an acoustic discriminability score which generates the probability that two tokens within a category are less distant than any two tokens in different categories. They also measured the phonological density of IDS and ADS, calculating the proportion of changes to be performed to transform one word to another (e.g., a minimal pair would have one segmental change. Acoustically, IDS was more variable and less distinguishable than ADS. Phonologically, IDS had more distinctive word-forms than ADS (likely due to the inclusion of onomatopoeia), but this phonological separation did not compensate for the acoustic level differences.

Beyond extensive corpus work on the phonetic variability of IDS, computational modelling provides a useful avenue for testing theoretical hypotheses of phonological acquisition, such as methods of phoneme identification. Ludusan et al. (2020) compared recordings of Japanese IDS, ADS, and read speech from a set of mothers and found modest evidence of hyperarticulation (measured as increased between-category distance) in IDS compared to ADS. The strongest

hyperarticulation, however, was observed in read speech. Further, they found that categories in IDS were not more separable (defined as both hyperarticulated and less variable) than ADS. In fact, they found worse generalization of classification of vowel categories when six different machine learning algorithms were trained on IDS compared to ADS. Again, read speech yielded the most robust data for vowel classification. As such, IDS not the best learning tool for phonemic categories, although it may serve an important purpose with regards to other aspects of language development.

Overall, these corpus and computational investigations of phonetic variability in IDS have predominantly approached phonetic variants as "noise" to be filtered out from the phoneme "signal", thus treating hyperarticulation as a beneficial modification of IDS. But phonetic variation poses an interesting learning challenge beyond the simple "noise filtration" approaches assumed in the previous characterizations of IDS outlined above.

1.3 Positional allophony

The previous literature concerning phonetic variation in IDS has overwhelmingly centered around the challenge of extracting the canonical form from variable language input. As a result, these studies focused mainly on comparing the extent of variation in IDS to ADS, since one possible solution to the phoneme extraction challenge would be that the input, i.e. IDS, contains fewer allophonic variants than ADS.

But extracting the underlying phonemic form isn't the only learning problem presented by phonological acquisition: as discussed in §1.1, not all phonetic variation is random. In fact, some phonetic variation is required and is precisely what must be learned, such as positional allophony. Thus, we must consider the phonological learning problem as two-fold: in addition to identifying

the canonical variant from a set, infants must also uncover when and where phonetic variants surface, while distinguishing them from unintentional results of reduction and assimilation at high rates of speech. Because positional allophony information is language-specific, infants can only use their IDS input to discover it.

Since so much of the previous literature on phonetic variation in IDS has focused only on the challenge of extracting the canonical form (and therefore sought to compare degrees of variation between ADS and IDS), we know relatively little about the respective timing of learning phonemes and learning positional allophony.

1.3.1 Acquisition of positional allophony

Extant development research shows that infants are sensitive to allophones early in life and are also able to use positional allophony cues to aid in word segmentation. English-learning 2-month-olds are able to discriminate between the /t/ in "night rate" [naɪʔ reɪt] versus "nitrate" [naɪtʃreɪt] (Hohne & Juscyk 1994). By 10.5 months-old, they are even able to use this allophonic variation to segment words (Juscyk et al. 1999).

However, the ability to use allophones to segment a speech stream isn't necessarily indicative of true *allophonic knowledge* – that is to say, a categorical understanding of which allophones correspond to which phonemes in the target language, and the licensing environments for each of these allophones. By 11-months of age, sensitivity to allophones appears to decrease. Siedl et al. (2009) trained English- and Quebec French-learning infants on a pattern that depended on vowel nasality (something which is allophonic in English but phonemic in French), then tested the infants' abilities to generalize this pattern to new syllables. Although English-learning 4 month-olds succeeded in this generalization, mirroring the performance of French-learning 11

month-olds, English-learning 11 month-olds no longer were able to learn this pattern and failed to generalize this phonotactic regularity. This suggests that infants' early representation of sound categories are phonemic, as opposed to allophonic,. But exactly how manageable is the task of learning phonemes from a set of allophones?

Again, computational models provide one plausible mechanism by which infants could learn allophones. Peperkamp et al. (2006) created a statistical learning algorithm which could discover phonemes and allophones based on complementary distributions from semiphonetically-transcribed French IDS. They found that the algorithm could only succeed if given pre-specified articulatory or perceptual features and assumptions regarding the nature of possible allophonic rules (e.g., the default segment and allophone must be neighbors in phonetic space). The performance of this algorithm was further improved upon inclusion of minimal pairs, suggesting word forms in IDS could be more beneficial for learning phonemic categories than purely distributional information (Martin et al. 2013). However, once again, a common thread within this corpus and computational infant allophone literature is that allophones are largely treated as "noise" that can either hinder or assist the acquisition of phonemic categories.

Before we can determine the acquisition trajectories of the two main phonological learning problems – extracting the canonical/phonemic form from a set of variants, and learning the positional allophony of the target language – we must first quantify the full extent of variability in the everyday speech directed to infants, as well as the positional predictability of this variation in the IDS input.

1.4 Approach

In this study, we evaluate how variable and predictable the most common segments of English are in everyday speech directed to infants. To do this, we transcribed ~7,000 utterances from the Providence Corpus (Demuth et al. 2006) to quantify the degree of variation present in coronals, some of the most variable consonants in English. In doing so, we have compiled one of the largest phonetically transcribed utterances to date. This is particularly important, as documenting the extent of allophonic variation in naturalistic IDS is critical in order to make future theoretical and computational modeling of phonological acquisition ecologically valid. Further, we determined the distribution of canonical and non-canonical (variant) forms of the coronal segments in naturalistic IDS, the contexts in which they are observed most often, and how predictable this positional variation is. From this analysis, we can begin to chart how and what infants can learn about positional allophony from their IDS input.

2 | Methods

2.1 Coding

2.1.1 Data

For this study, we analyzed the Providence Corpus (Demuth et al. 2006) to determine the rate of occurrence of canonical and non-canonical variants of coronals present in naturalistic infant-directed speech (IDS). This longitudinal corpus consists of home audio and video recordings of parent-child interactions in 5 monolingual English-learning children during everyday activities. We sampled these recordings at two ages for our phonetic analysis: 16-18 months and 22-24 months.

First, we identified each mother's utterances containing the target segments. This was done by extracting any utterances in the orthographic transcript that were coded as the mother's utterances and contained "t", "d", "s", "z", or "n", since the orthographic symbols of these segments correspond almost exclusively to their phonetic equivalents. The time points for each utterance were then used to extract the relevant portion of the recording, which was then force-aligned using the Forced Alignment and Vowel Extraction program suite (FAVE; Rosenfelder et al. 2014). This force-aligner uses an HTK Toolkit for phonetic alignment, referencing the CMU Pronouncing Dictionary to transform orthographic transcription into phonemic notation. Altogether, this yields a set of Praat (Boersma and Weenink 2013) TextGrids containing a time-aligned segmented phone (phonemic) tier and a word tier. Any segments on which FAVE failed (because the words were not in the pronunciation dictionary) were excluded and not annotated (6,586 tokens or 1.9%).

2.1.2 Annotations

The boundaries of the relevant force-aligned segments were then checked, realigned, and annotated by three phonetically-trained research assistants, all native speakers of American English. These segments were annotated for the phonemic form (based on the FAVE output), the surface form (the phonetic variant that surfaced), word position (initial, medial, or final), and surrounding segments. The possible surface variants for each underlying form are shown in *Table 1*. Cross-coder reliability was 74.13%. Annotation criteria for each of the variants coded can be found in Appendix A, and representative spectrograms for each variant can be found in Appendix B.

Table 1. Possible Surface Variants of Coronal consonants.

Phoneme	***Surface Variants***
/t/	[tʰ] (aspirated), [t] (unaspirated), [t̬] (voiced), [ɾ] (tap), [ʔ] (glottal stop), [t̚] (unreleased), [t̪, tʷ, tʸ] (assimilated), [tʃ] (affricated), [Ø] (deleted)
/d/	[d] (canonical), [ɾ] (tap), [d̥] (voiceless), [d̚] (unreleased), [dʒ] (affricate), [d̪, dʷ, dʸ] (assimilated), [Ø] (deleted)
/n/	[n] (canonical), [ɾ̃] (nasalized tap), [n̪, nʷ, nʸ] (assimilated), [n̩] (syllabic), [Ø] (deleted)
/s/	[s] (canonical), [s̬] (voiced), [ʃ] (assimilated), [Ø] (deleted)
/z/	[z] (canonical), [z̥] (voiceless), [ʒ] (assimilated), [Ø] (deleted)

2.1.3 Exclusions

Because we were interested only in naturalistic IDS, tokens were excluded from the analysis for the following reasons: mechanical/acoustic noise (such as microphone static or feedback); obviously adult-directed speech; reading or singing; child vocalizations and speech; and routinized expressions, such as *"wanna"* or *"gonna"*. Additionally, because the files were sampled using the corresponding orthographic symbols for each segment, some number of files were sampled that did *not* contain any of the target segments – e.g., an orthographic "t" could actually correspond to [θ], leading the file to be sampled, but ultimately excluded if there were no coronal segments. A total of 7,056 utterances were excluded from the original transcripts for these combined reasons. The number of utterances analyzed in the final sample was 6,750, with 33,335 tokens annotated. While it may initially seem alarming that so many utterances were excluded from the analysis, it is expected that a significant portion of the recorded input will be "overheard" speech between adult caretakers (i.e. ADS), because these are at-home, naturalistic recordings (see Shneidman & Goldin-Meadow 2012, Shneidman et al. 2013, and Weisleder & Fernald 2013 for discussions of the role of overheard ADS in language acquisition).

2.2 Analysis

Once the corpus annotation was complete, the annotated segments were extracted using a custom-written Python program. A small number of tokens (1,083) were excluded due to a script error or contained missing data (i.e. missing realization, position in word, preceding segment, and/or following segment). The total number of analyzed segments was 31,245.

3 | Results

3.1 Overall variability in IDS

Cross-linguistically, syllable onsets are more faithful to their underlying form than codas (e.g., Beckman 1998) – by extension, word-initial consonants are expected to be more faithful/canonical than word-final consonants. The "canonical" form of /t/ is the subject of some theoretical debate. While it is generally accepted that canonical /t/ exhibits a distinct stop closure and release, there are mixed views on the underlying specification with regards to aspiration. Some have argued that English voiceless stops are specified as underlyingly [-aspirated] (Odden 2005), others adopt released, [+aspirated] as canonical (Vaux 2002), and still others do not specify, treating [±aspirated] as canonical (Dilley et al. 2019). It is this last convention that we will adopt in discussions of canonical /t/ in this paper. Elsewhere, we will separate released, unaspirated [t] from aspirated [t^h] with no claims on canonicality. The distribution of canonical forms for each segment across word positions can be found in *Table 2*, below.

A logistic mixed effects model with a random intercept for subject, and a fixed effect of position (initial vs final), confirmed that every segment had more canonical instances in initial compared to final position (p's < 0.002). However, as we can see there is a substantial difference in the extent to which individual segments are canonical; across all word positions, canonical

variants of /t/ are in the minority (39.2%) in comparison to /s/, which is overwhelmingly canonical (98%).

Table 2: Distribution of canonical forms (raw count | percent).

Sound	*Initial*	*Medial*	*Final*	*Overall*
/t/	1456 | 91.5%	1599 | 62.3%	1000 | 20.6%	4055 | 39.0%
/d/	1508 | 82.6%	406 | 40.9%	729 | 31.1%	2643 | 51.2%
/z/	54 | 68.4%	208 | 88.5%	2701 | 75.6%	2963 | 76.2%
/n/	1202 | 98.6%	2633 | 88.9%	2876 | 93.4%	6711 | 92.5%
/s/	2491 | 98.8%	749 | 98.7%	3057 | 97.0%	6288 | 97.9%

In Figure 1, we show the details of the phonetic instantiation of segments by position and variant type. For ease of comparison, all variants are listed for all segments, despite the fact that

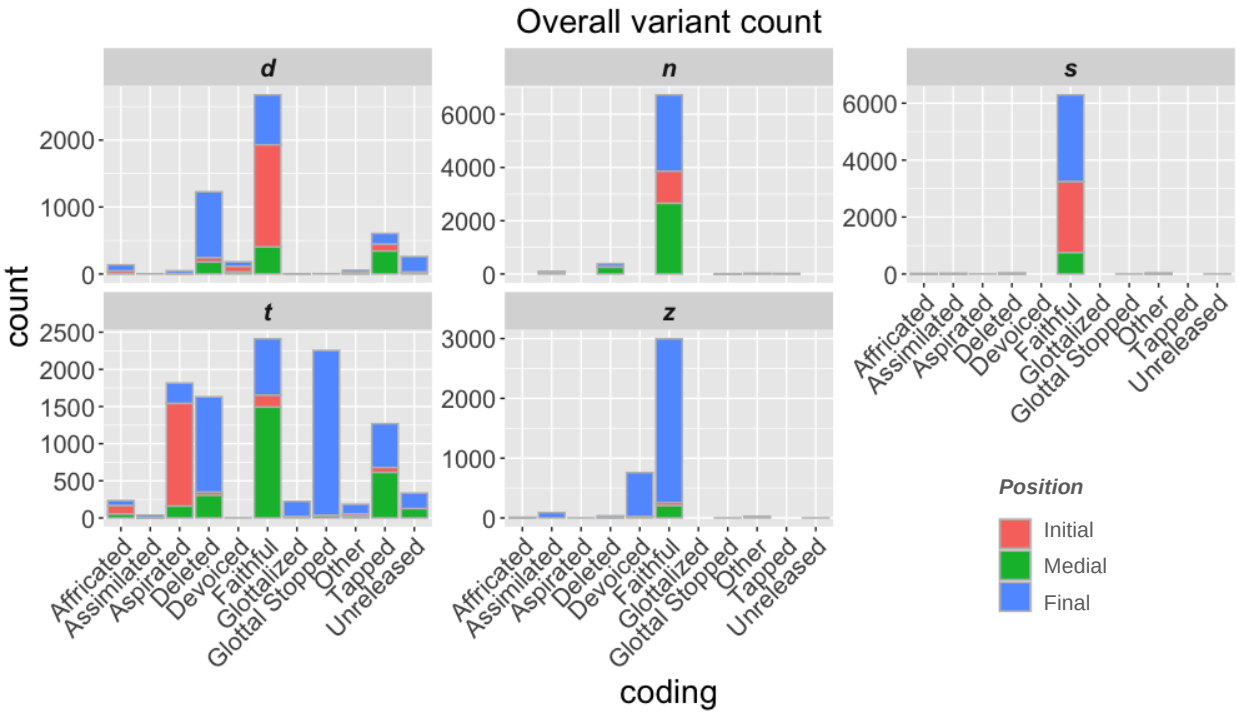


Figure 1: Surface variant counts (by position) for each segment.

the expected (and observed) frequencies of some of these cells are 0. For example, we do not expect /s/ and /z/ to be tapped in any environments, and this is what we observe – the frequencies of tapped /s/ and /z/ are both 0. Note also that "Faithful" in Figure 1 denotes those variants that surfaced as the phonemic form – that is, instances where /t/ surfaces as released, unaspirated [t], /d/ surfaces as released [d], etc. As previously discussed, this is not always the *canonical* variant.

Clearly, the distribution of variants differs from category to category – as expected, based on the number of possible allophones for the different categories (Table 1). While /s/ is almost entirely faithful, /t/ has highly frequent variants in addition to phonemic [t]– indeed, there are almost as many glottal stops word-finally as there are [t] across all word positions.

3.1.1 Variation in IDS compared to ADS

We observe the following overall trends in our data: /t/ is the most frequently occurring category (10,398 segments), followed by /n/ (7,267), /s/ (6,425), /d/ (5,205), and /z/ (3,952). This is consistent with previous reports: in American English naturalistic ADS, /t/ is the most frequent consonant and comprises ~8.4% of all phonemes in spoken American English (Denes 1963). The next most frequent sound is /n/ (7.1%), followed by /s/ (5.1%), /d/ (4.2%), and /z/ (2.5%). Thus, the distribution of segments in our data mirrors the general trends observed in ADS.

However, it is important to note that although much of the past literature on IDS – as discussed §1.2.2 – has centered around the relative proportions of canonical forms in IDS compared to ADS, it is higher exposure to IDS, and *not* ADS, that is correlated with larger vocabularies and faster lexical processing in infants (Bergelson et al. 2019, Huttenlocher et al. 2010, among others). Because infants seem to learn primarily from IDS (not ADS), in this study

we shift away from any further comparison to ADS, focusing instead on the distributional information alone. This more accurately sets up the context for phonological learning for infants, as they are not privy to the distribution of variants in ADS and must instead learn from the input with which they are presented.

The following sections break down the distribution of variants in the coronal segments analyzed. Note that in these sections, we only list and discuss the distribution of the three most frequent variants for each of the segments, unless otherwise stated. The full raw count distributions of the observed variants can be found in Appendix C.

3.1.2 Alveolar plosives (t and d)

In English, /t/ and /d/ are among the most frequent sound categories, comprising over 12% of total speech sounds in ADS (Denes 1963, Tobias 1959, Zue and Laferriere 1979). Across all word positions, we observed 10,398 instances of /t/, of which the three most common variants were [t] (2,441), [ʔ] (2,255), and [t^h] (1,818). Of 5,205 instances of /d/, the three most common variants were [d] (2,675), [Ø] (1,228), and [ɾ] (606). Thus, [t] and [ʔ] are equally frequent instantiations of /t/ in English IDS across all positions, but /d/ is most likely to be produced as [d].

A different picture emerges if we consider word-position. In word-initial position, of 1,797 instances of /t/, 1,382 surfaced as aspirated [t^h] as expected, with 161 surfacing as [t] and 113 surfacing as [tʃ]. Of 1,839 target /d/, in word-initial position 1,526 surfaced as [d], 102 as [ɾ], and 84 as [d̥]. For /t/, the canonical variant is [t/t^h] and for /d/ the canonical variant is [d]; these are the most frequent variants we observe, comprising 91.5% and 83% of the word-initial targets, respectively (Table 2).

In word-medial position, of 5,718 instances of /t/, 1491 surfaced as [t], 612 surfaced as [ɾ] and 304 as [ø]. Of 987 instances of /d/, 403 surfaced as [d], 342 surfaced as [ɾ] and 177 as [ø]. In medial position, many /t/ and /d/s were tapped intervocalically. In ADS, taps comprise 76 to 99% of the word-medial variants of these categories in licensed conditions (Herd et al. 2010, Patterson and Connine 2001, Zue and Laferriere 1979). In our data as well, taps surfaced in 75.4% of licensed positions (954 taps observed in 1,265 environments).

In word-final position, of 5,782 instances of /t/, the three most common variants were [ʔ] (2,222), [ø] (1,285), and [t] (759). Of 2,379 instances of /d/, 982 were deleted, 746 surfaced as [d], and 233 surfaced as [d̚]. That is, word-finally, /t/ is mostly produced as a glottal stop, three times

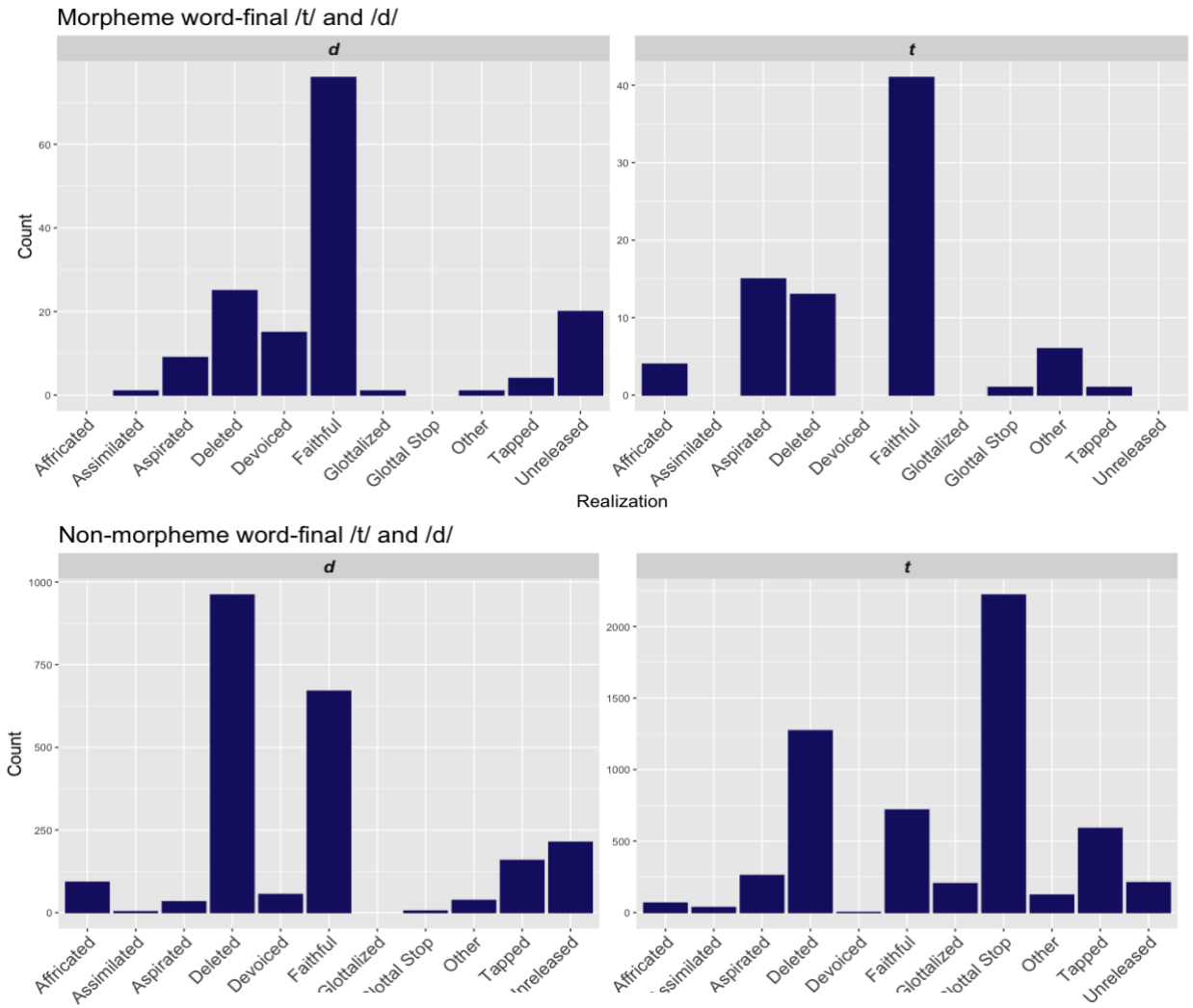


Figure 2: morphological and non-morphological /t/ and /d/

more often than as a canonical variant, whereas /d/ is deleted more often than produced canonically.

Lastly, we consider the distribution of variants as a function of morpheme. In English, regular past-tense morphology is marked by the word-final morpheme *-ed*, which can be instantiated with a word final /-t/[1] or /-d/. To determine any differences in variant distribution between morphologically conditioned /t/ and /d/ and other word-final instances /t/ and /d/, we filtered these underlying forms in regular past-tense words (with a word-final *"-ed"* in the target).

The distribution of variants in morphological and non-morphological (words like *"spot"*) word-final /t/ and /d/ is shown in Figure 2, above. Suffix /-t/ and /-d/ are predominantly faithful to the phonemic form (Figure 2, top) in contrast to words ending in /t/ and /d/, which are overwhelmingly non-canonical, with the most frequent surfacing form for /d/ being a deletion and the most frequent surfacing form for /t/ a glottal stop (Figure 2, bottom).

3.1.3 Alveolar nasal (n)

Across all word positions, we observed 7,267 /n/. Of these, the two most frequent surfacing variants were [n] (6,725) and [ø] (389) – thus, /n/ is predominantly faithful. In word-initial position, /n/ surfaced faithfully ([n]) 1,206 times out of 1,222 total word-initial /n/s (98.7% canonical). Out of 2,985 word-medial /n/s, 2,651 surfaced as [n] and 252 as [ø] (88% canonical). Interestingly, very few word-medial /n/s surfaced as taps – only 10 instances of taps were observed word-medially. Word-finally, 2,868 surfaced as [n], 128 were deleted, and only 37 assimilated out of 3,060 word-final /n/s (93.7% canonical).

[1] The established analysis for the past-tense suffix is that the underlying form is /-d/. Here, we treat the expected surface [-t] as phonemic only for ease of comparison between morphological and non-morphological word-endings. In reality, the past-tense [-t] is derived from underlying /-d/, and is not phonemic.

3.1.4 Alveolar fricatives (s and z)

Across all word positions, we observed 6,425 /s/ and 3,952 /z/. For /s/, the two most frequent surface variants were [s] (6,292) and [∅] (41) – thus, /s/ is predominantly faithful. Overall, /z/ was slightly less canonical, with 2,996 surfacing canonically ([z]) and 758 devoiced ([z̥]).

In word-initial position, /s/ surfaced canonically ([s]) 2,491 times out of 2,513 total word-initial /s/s (99.1% canonical). /z/ in word-initial position is comparatively rare (55 instances); all 53 surfaced canonically. Out of 760 word-medial /s/, 750 surfaced as [s] (98.7% canonical). Word-medially, of 231 /z/s, the two most common variants were 205 surfacing as [z] and 23 devoiced [z̥].

Out of 3,152 word-final /s/, 3,051 surfaced as [s] while 33 as [∅] (96.8% canonical). Of 3,639 word-final /z/, 2,996 surfaced faithfully ([z]) and 758 were devoiced [z̥] (82.3% canonical). As with /t/ and /d/, it is important to note that word-final [s] and [z] are used in English to mark morphology - regular plural s ("books"), possessive -s ("Sarah's"), third-person singular -s ("walks"), and clitic -s (contractions of "has" and "is"). In our dataset, morphemic /t/ and /d/ comprise only 2.9% of word-final instances. In a sample of the Buckeye Corpus (Pitt et al. 2005), Plag, Homann and Kunter (2017) found that non-morphemic /s/ and /z/ comprised 30.8% of all word-final instances of /s/ and /z/. Although separating out morphemic /s/ and /z/ for analysis is beyond the scope of this study, it is worth noting that because we observe little variation in /s/ and /z/ (average 89.6% canonical forms for /s/ and /z/, compared to average 21.95% for /t/ and /d/) it is unlikely we will observe a significant difference in suffix *-s* compared to other word-final /s/.

3.1.5 Summary

Taken together, the distribution of phonetic variants for each segment showcases several patterns. First, across all segments, the phonemic/canonical variant is the most frequent variant observed. Indeed, in all cases but /t/, the canonical variant comprises the majority (i.e. greater than 50%) of the surfacing forms for each segment. Second, the distribution of variants differs from segment to segment. While /s/ and /z/ only have 4 possible allophones (Table 1) and overwhelmingly surface canonically (92.5% and 97.9%, respectively), alveolar plosives /t/ and /d/ have the lowest proportion of canonical forms (39.0% and 51.2% respectively). This is expected given the overall higher number of possible allophones (Table 1) and is reflected in the frequency of non-canonical variants (Figure 1).

But the higher number of theoretically expected variants cannot explain the entirety of the observed variation in IDS – for example, despite having a similar number of expected variants, the distributions of /t/ and /d/ are distinct, with /t/ having at least 3 highly frequent variants. In the following sections, we will turn our attention to how much of this variability across segments can be explained using phonological context and theorized rules of positional allophony.

3.2 Context predictability of variants

As discussed in §1.3, extracting the canonical form from a set of variants is not the only challenge that young phonological learners face. Another crucial aspect of phonological acquisition is learning how and when particular positional allophones must surface. To determine the quality of distributional information available in the IDS input, we begin by investigating how predictable the variation observed in this data set is. We do this in two ways: first, we calculate the overall predictability for each segment using the information-theoretic measure of entropy.

Second, we generate a preliminary analysis of expected allophones using a set of simplified positional allophony rules.

3.2.1 Entropy

One measure of overall event predictability is entropy. Entropy is a metric of the predictability of any outcome (Shannon 1948): in the case of a particular event for which there is only one outcome, the entropy for that event would be 0. The formula for entropy is:

$$H(X) = -\sum_{i=1}^{n} P(x_i)\, log_2\, P(x_i)$$

where x_i is one of the possible outcomes for an event X, and $P(x_i)$ is the probability of that outcome. Higher values of entropy indicate greater uncertainty in the outcome (and, conversely, lower predictability). In our case, the higher the entropy for a particular sound category, the less predictable the surfacing variant for that sound. This measure allows us to compare the overall predictability for each category despite differences in overall frequencies. Furthermore, highly

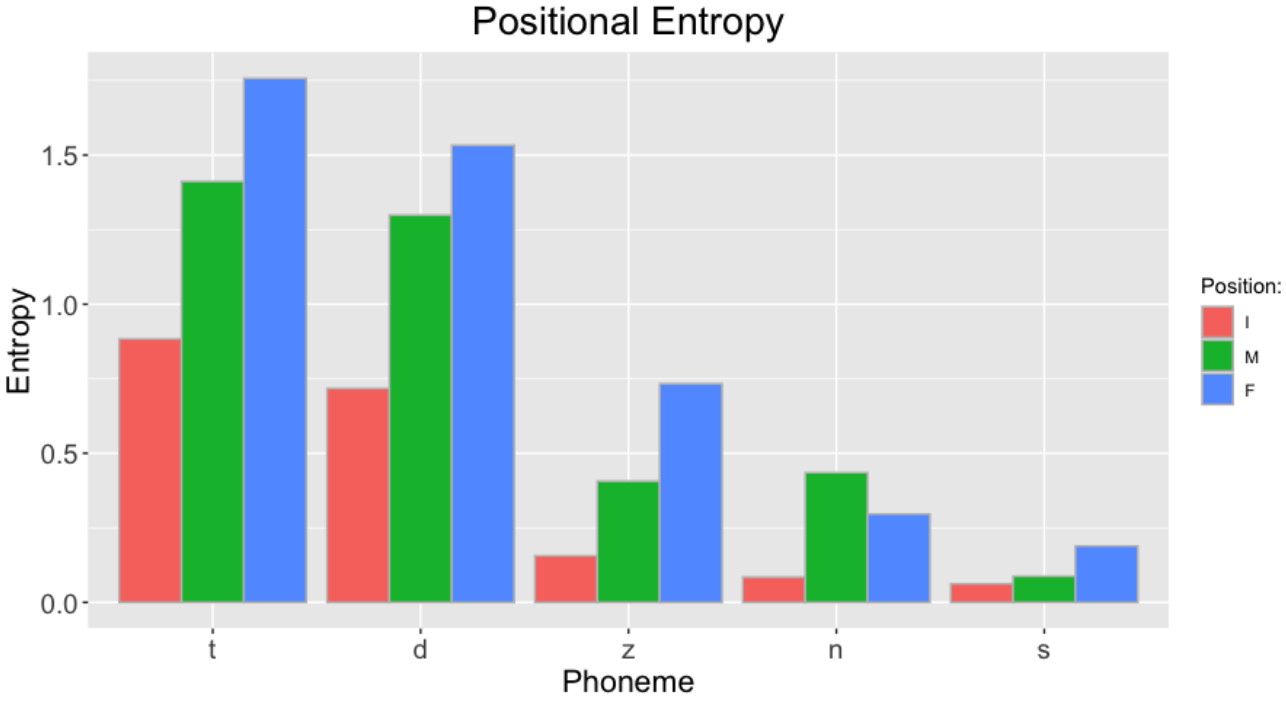


Figure 3: Entropy of each phoneme by position in word.

entropic (and therefore highly unpredictable) categories could be expected to pose greater challenges to the phonological learner. The entropy for each target sound, further divided by position in word, is shown in Figure 3.

Because entropy increases as function of the number of possible outcomes, it is always higher for a segment that has more surface variants. Therefore, /t/ has the highest overall entropy, indicating a greater number of variant surface forms, followed by /d/. As expected, /z/, /n/, and /s/ all have lower overall category entropies. These results are expected given the overall larger number of possible surface variants of /t/ and /d/ in English listed in Table 1.

Additionally, by comparing entropy for a segment across positions, we can see that entropy in initial position is systematically lower for every segment. Thus, the initial position is privileged not only because it has a higher proportion of canonical instances, but also because the surface variation there is more predictable.

Finally, despite having 3 surface variants each, there is a difference in entropy for /s/ vs. /z/; specifically, /z/ has a higher positional entropy compared to /s/ – indicating that surface variants for /z/ are less predictable. Additionally, recall that /z/ also has a lower proportion of canonical instances, presenting a greater challenge to a phonological learner. A similar asymmetry is observed with word-final /t/ and /d/, albeit in the opposite direction: it is /t/ that has higher entropy than /d/, indicating that word-final /t/ is less predictable than /d/. Similarly to /z/, /t/ also has a lower proportion of canonical instances than word final /d/, and is thus expected to pose an additional challenge to the phonological learner.

Next, we examine whether the rank ordering of entropy in subject to individual variation. The entropy of each phoneme, divided by speaker, is shown in Figure 4, below. As we can see,

although the absolute value of entropy varies across speakers, the relative differences in entropy across position and segments remain consistent between them.

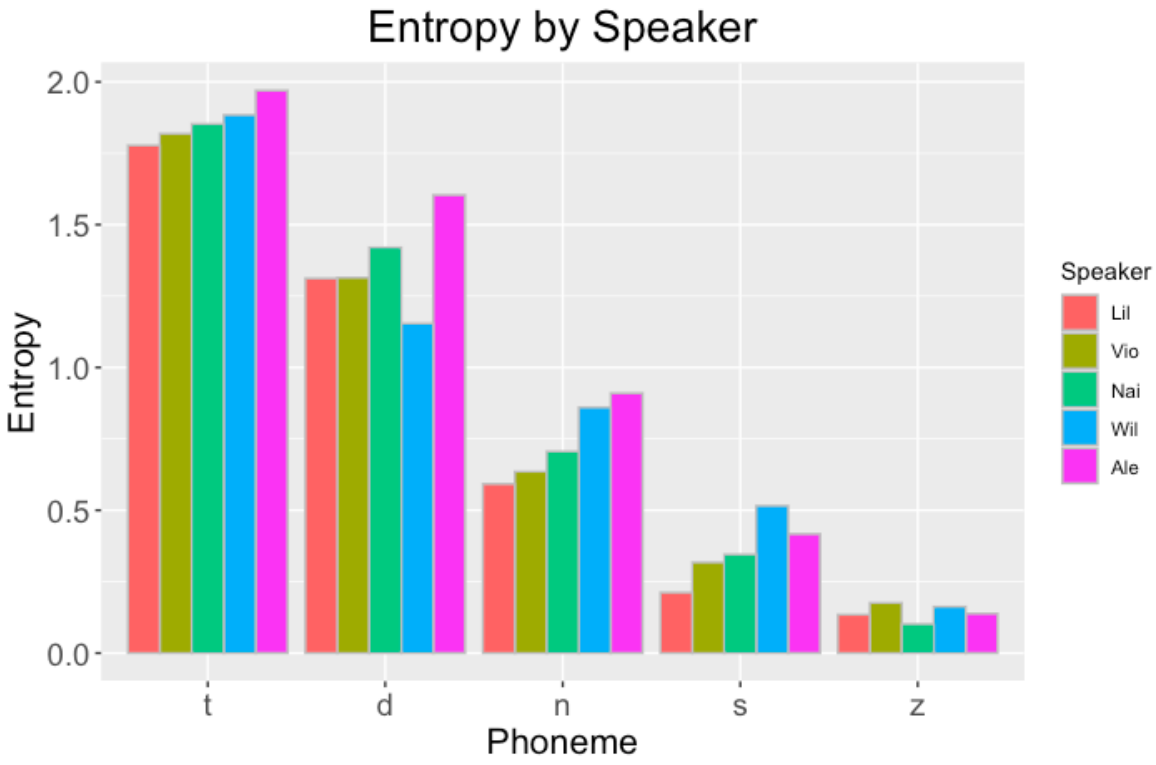


Figure 4: Entropy of each phoneme by speaker.

We observe no consistent outliers across the five speakers, which suggests that the entropy distributions shown in Figure 3 cannot be attributed to any one, highly variable speaker.

3.2.2 Confusion Matrix

As previously discussed in §1.3, some degree of positional variation is governed by phonotactic and phonological rules. While generating expected allophones based on phonological environment would require a comprehensive phonetic grammar of English (which has not yet been established), we can begin to approximate the degree of expected and observed variants based on some theorized rules of English phonetics (Ladefoged and Johnson 2014). For this reason, the following section is intended primarily as a proof of concept. This is an important first step in

quantifying how predictable the distributional evidence for positional allophony is in the IDS input children receive. Future work will involve a more nuanced, gradient approach to generating the expected allophones for the various phonological environments observed in American English.

In order to determine the relative frequencies of expected variants and observed variants, we generated a confusion matrix obtained by plotting expected versus observed variants for /t/ and /d/ (Figure 4), two segments that are the most variable. A list of the rules used to generate predicted/expected variants can be found in Appendix D.

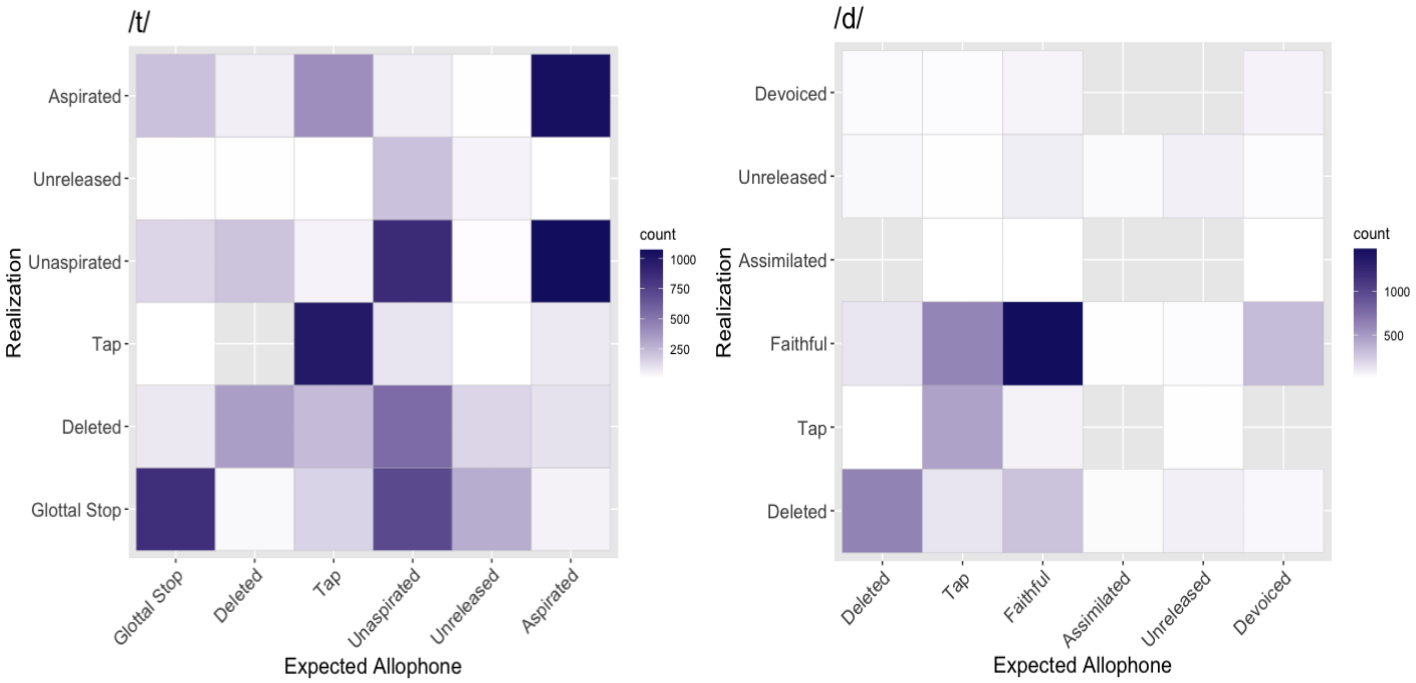


Figure 5: Expected vs. realized variants for /t/ (left) and /d/ (right). Grey squares without color denote combinations that were not observed.

A match between the expected and realized variant is indicated by the highest (dark blue) counts on the diagonal; however, we see that unexpected variants (light blue/purple cells outside the diagonal) are common. Overall, for almost every expected variant there was at least one other unexpected variant that was highly likely.

For /t/ (Figure 5, left), when aspirated [tʰ] was expected, [t] was approximately just as frequent as [tʰ]. When unaspirated, released [t] was expected, however, several other highly frequent variants were also observed, including [ʔ], deletion, and [t̚]. In tapping environments, although the realized tap was most frequent variant, [tʰ] also surfaced. When a glottal stop was expected based on phonological environment, [tʰ] was also frequent. For expected deletions, [t] was almost as frequent as the deletions. Interestingly, in the case of expected unreleased [t̚] (before another stop), a [ʔ] or deletion was even more frequent than the expected variant.

For /d/ (Figure 5, right) canonical [d] surfaced most often, although they were occasionally deleted. Medial expected taps were evenly divided between taps and [d]s (although this could be because these are difficult to distinguish perceptually; see de Jong 1998, Herd et al. 2010, Malécot & Lloyd 1968). When a deletion was expected, this was the most frequent variant. Assimilated and unreleased /d/ were highly infrequent or non-existent. Lastly, for expected [d̥], the most frequent realized variant was the canonical [d].

Overall, across all expected allophones, the most frequent variants for /t/ and /d/ were [t, ʔ, ɾ, tʰ] and [d], respectively. For /t/, in both the glottal stop and tapping environments, the expected variant far exceeded the canonical variant. The unaspirated ("faithful") variant was equally as frequent as the expected variant in aspirated contexts. In the case of expected faithful variant [t], an unexpected [ʔ] is almost as frequent. For /d/, the canonical variant exceeded the frequency of expected variants in tapping and devoicing environments; the expected deletions were more frequent than the canonical variant in deletion environments.

Although frequency is an important factor in distributional learning of positional allophones, we also need to consider the relative predictability of each expected variant. To determine the predictability of variants in a given environment, we generated an entropy plot for

each variant of /t/ and /d/ in its respective licensing environment (Figure 6). If, for example, only taps ever surfaced in the tap licensing environment, the entropy for that environment would be 0.

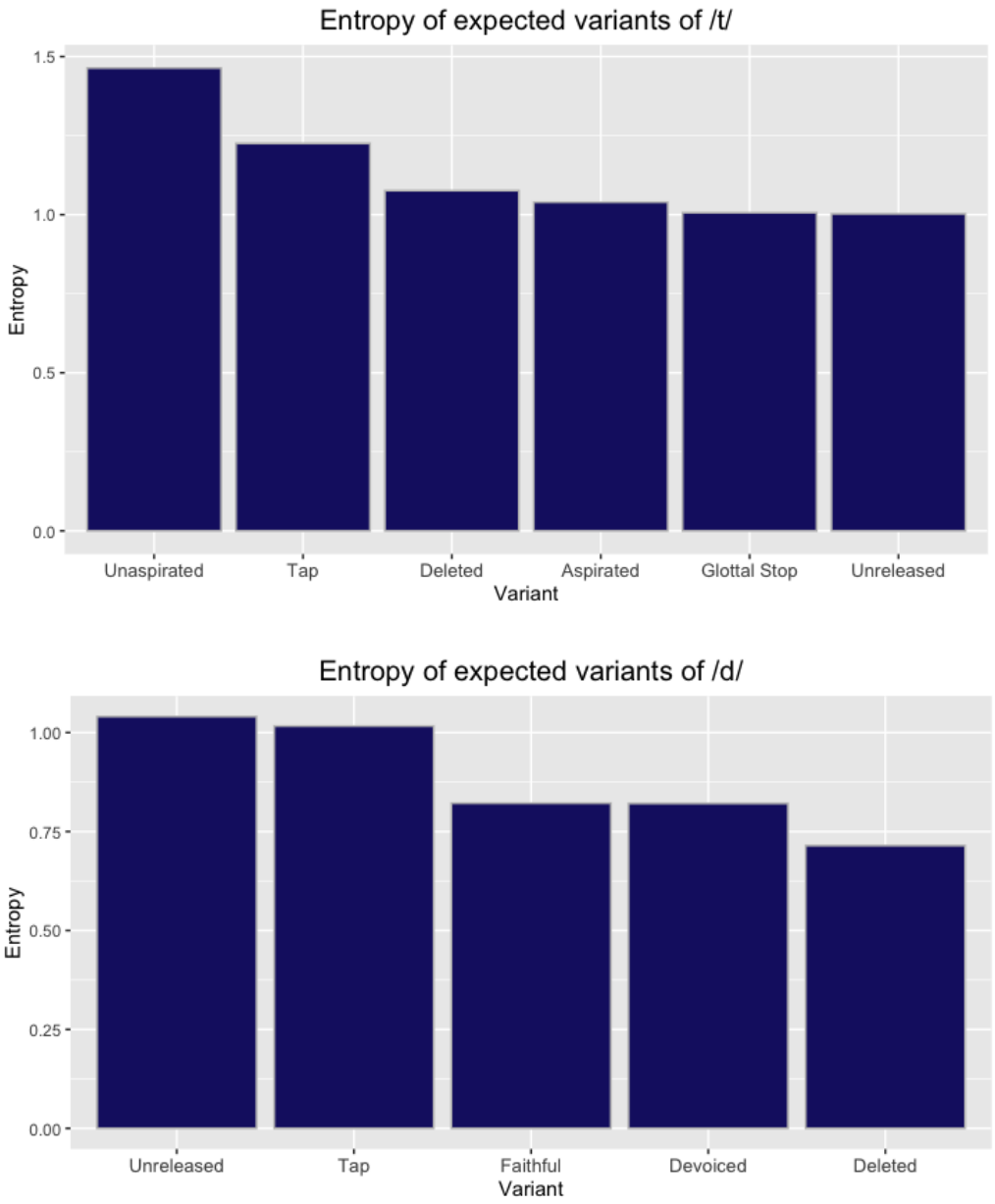


Figure 6: Entropy of surfacing variants within the expected variant environments[2].

For /t/ (Figure 6, top), expected [t] has the highest entropy at 1.46, followed by taps (1.23). The remaining variants all had entropies that roughly fell around 1: [ʔ] (1.01), [tʰ] (1.04), deleted

[2] Assimilation was not included in these entropy calculations due to the presence of more than one expected variant.

(1.07) and unreleased variants (1.00). For /d/ (Figure 6, bottom), the expected unreleased variant had an entropy of 1.04 while expected [ɾ] had an entropy of 1.02. Devoiced [d̥] and canonical [d] had equal entropies of 0.82; deletions were most predictable with an entropy of 0.71.

When it comes to learning positional allophony, higher entropy variants would indicate "noisier" data for the variant – even if a variant is highly frequent in a particular environment, the overall unpredictability (high entropy) of a variant in its respective environment could pose additional challenges to the learner. These challenges are discussed further in §4.2 and §4.4.

4 | Discussion

In this study, we evaluated the extent of variation observed in the pronunciation of speech sound categories in spontaneous IDS in order to provide a holistic view of the phonological learning problem. To do this, we phonetically transcribed ~31,000 segments from the Providence Corpus (Demuth et al. 2006) to quantify the degree of variation present in coronals, some of the most frequent and variable consonants in English. We then calculated the proportion of canonical and non-canonical (variant) forms in IDS and the contexts in which they are observed most often. Lastly, we performed a preliminary analysis to determine how much of this variation can be captured by a set of theorized positional allophony rules.

Taken together, our analyses of phonetic variation in IDS highlight several clear patterns: first, the canonical variant for a sound category is typically the most frequent variant produced in IDS. Second, the expected variant based on the surrounding phonological environment is not always the most frequent variant to surface. Lastly, morphologically-conditioned word-final consonants are more faithful than those that are not morphologically conditioned.

4.1 Canonical variants are typically most frequent

Our findings indicate that the canonical variant is typically the most frequent surface variant (Table 1, Figure 1). This suggests that by attending to the most frequent variant in the IDS input, it may be possible to learn the underlying (and therefore canonical) form based only on distributional information. Indeed, if we consider the learning problem of extracting the abstract underlying phoneme from the variants that are present in the child's input, we could predict that based on the frequency of variants (and general predictability of each sound category, Figure 2) that the order of acquisition would range from the most faithful/canonical to the least canonical. That is to say, we would expect children acquiring American English to learn /s/ and /z/ before /n/, /t/, and /d/.

4.2 Variation cannot be fully explained by phonological environment

Using the surrounding segments for our target sounds, we performed a preliminary positional allophony analysis by generating the expected allophone based on English phonotactics (Ladefoged and Johnson 2014). We then calculated the counts of expected variants compared to the actual realized surface variants (Figure 5). Entirely context-predictable variants would be indicated by a clear diagonal, where the highest counts occur where realized and expected variants match. However, our findings indicate that for each variant of /t/, there is at least one other highly frequent variant. This suggests that learning positionally-governed phonetic variants is complicated by the presence of noise, such as reduction and assimilation processes produced at high rates of speech. Furthermore, in some cases, an unexpected variant is more frequent than the expected variant (e.g. [ʔ] instead of word-final [t]).

Additionally, the overall predictability of expected variants differs from variant to variant. For example, although [ɾ] is more frequent in tapping environments than is [ʔ] in glottal stop environments, [ʔ] has lower entropy and is thus more predictable than [ɾ]. If frequency alone determines order of acquisition, we would expect [ɾ] to be acquired before [ʔ]. If, on the other hand, the predictability of a surface variant is more important for acquisition, we would expect infants to learn about the [ʔ] before [ɾ]. These "noisy but frequent" surfacing variants may therefore pose an interesting opportunity to determine the constraints governing when infants acquire positional allophony.

4.3 Morphological variants are more faithful

Lastly, our findings indicate that while the morphological word-final /-t/ and /-d/ are predominantly faithful, non-morphological word-final /t/ and /d/ are *not* – word-final /t/ is more likely to surface as a glottal stop, while /d/ is more likely to be deleted. There are many fewer regular past-tense /t/ and /d/s than other target words containing word-final /t/ and /d/, and thus a direct comparison is not possible. However, the distribution is clearly distinct, as are the relative proportions. This indicates that the functional load of the segment also contributes to the extent of surface variation observed in IDS.

It is possible that the observed difference in variant distribution across morphological and other word-final /t/ and /d/ may be affected by factors not considered in this study: specifically, perhaps by virtue of being a past tense marker, *-ed* may occur in a subset of all possible phonological environments. For example, the past tense of "play" would likely occur in constructions like "played with" – thus this *-ed* would occur more frequently with the following segment /w/ in the phonological environment. Further analysis will be needed to determine

whether there are any significant differences in the distribution of phonological environments as a result of the morphological and syntactic patterns of English – and whether morphological conditioning further affects the distribution of these phonological environments and the observed variants.

4.4 General discussion

While previous work on IDS has largely centered around the problem of extracting the canonical form of a segment from the IDS input, the present study shifts away from this approach. Because phonological acquisition must go beyond identifying the canonical variants to the additional challenge of learning positional allophony, we instead aim to quantify the overall degree of phonetic variation present in naturalistic IDS *and* determine the context predictability of this variation. In doing so, we are able to gather a more holistic view of the day-to-day input infants receive and speculate on the implications of our findings for phonological acquisition, specifically the trajectory of learning phonemes and positional allophony.

In this corpus analysis, we found that the canonical variant for each of the coronal consonant sounds is always the most frequent surface variant overall, suggesting that canonical variants could theoretically be extracted from the input simply based on frequency. We found that positional allophones are also highly frequent, especially for /t/ and /d/. These findings have implications for the first of the two phonological acquisition problems approached in this study: identifying the phoneme from a set of phonetic variants. Based on this, we expect that extracting the canonical form from the distributional data would be more difficult in categories with higher degrees of positional variation (/t/ and /d/). Thus, under this view, /s/ and /z/ would be acquired first given their high rates of canonical variants across position and low entropy overall, and /n/,

/d/ and /t/ would be acquired later, as the phonemic form must be isolated and identified from a greater number of frequent surface variants.

With respect to the second acquisition problem, learning positional allophony, we can predict that some positional variants – for which there is clear and consistent evidence in the input – will be acquired before others. Specifically, we would expect taps [ɾ] and glottal stops in their respective phonotactic environments to be acquired before the aspirated [tʰ] or unaspirated [t] in their respective environments, given the higher frequency of positionally-appropriate variants surfacing in these positions. We would also expect [ʔ] to be acquired before [ɾ] due to its lower entropy and, therefore, higher predictability. Similarly, we would expect tapped [ɾ] and devoiced [d̥] to be acquired later than canonical [d], due to the higher frequency of faithful [d] in each of their respective phonotactic environments – with [d̥] acquired before [ɾ] due to its lower entropy and higher predictability.

Simply identifying the most frequent surfacing variant and the context predictability of these variants does not address whether infants have knowledge of positional allophony before, after or at the same age as they have knowledge of canonical variants. Our next step will be experimentally determining whether 12 month-olds prefer the canonical form or most frequent positional allophone in familiar and unfamiliar words. If infants are already aware of the rules that govern positional allophony, they should prefer the correct variant in the appropriate phonological environment. If, on the other hand, infants are only aware of the canonical form, we would expect them to prefer the canonical form, even when it is not in an expected position.

To date, computational models of language acquisition rely on idealized speech input, in part due to the current dearth of available IDS corpora that are fully transcribed. Idealized speech input significantly simplifies the acquisition problem and makes it less ecologically valid (see

Phillips & Pearl 2015, Feldman et al. 2013 for a more detailed discussion). In order to realistically model the problem of language acquisition, we need to start using speech input that closely resembles what children typically hear. This project is the start of a longer annotation project aiming to annotate all of the consonants in this IDS corpus. This will give us a holistic view of variation within IDS and generate the first quantified account of all consonants in naturalistic IDS. Beyond our analysis, our full, phonetically annotated corpus will be made available to the broader research community in order to fill this gap and allow for more accurate modeling of language acquisition.

5 | Conclusion

In this study, we set out to quantify the degree of phonetic variability in coronals in English IDS through a large-scale corpus analysis. We phonetically transcribed coronals in roughly 7,000 utterances of naturalistic IDS for their underlying form, surface form, position in word, and phonological environment. This data highlighted three salient patterns: 1) canonical variants for each sound category are (almost) always the most frequent; 2) expected variants, as governed by English phonotactics, often have at least one other highly frequent, unexpected variant; 3) morphologically-conditioned word-final alveolar stops are more canonical than content word-final alveolar stops. These findings suggest that although phonemes could potentially be acquired simply by "dialing in" to the most frequent surface variants from the variable input, learning positional allophony may be even more complicated than initially thought. In our analysis, even positional allophones do not follow entirely predictable patterns in the speech input to infants. Taken together, these findings have implications for computational models of phonological

acquisition given the high degree of phonetic variability observed in IDS, and the unpredictable patterns of observed allophones.

Bibliography

Beckman, J. (1998). Positional faithfulness (Doctoral dissertation). Amherst: University of Massachusetts, Amherst.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2), 1001–1024.

Bernstein Ratner, N. (1984). Patterns of vowel modification in mother–child speech. *Journal of Child Language*.

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3. 51. *Online: http://www. praat. org*.

Buckler, H., Goy, H., & Johnson, E. K. (2018). What infant-directed speech tells us about the development of compensation for assimilation. *Journal of Phonetics*, 66, 45–62.

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*(5572), 1435-1435.

Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., & Floccia, C. (2012). Linguistic processing of accented speech across the lifespan. *Frontiers in psychology*, 3, 479.

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. Journal of Child Language, 41(4), 913–934.

Dalby, J. M. (1986). Phonetic structure of fast speech in American English (Vol. 7). Reproduced by the Indiana University Linguistics Club.

De Jong, K. (1998). Stress-related variation in the articulation of coda alveolar stops: Flapping revisited. *Journal of Phonetics*, 26(3), 283-310.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, 49(2), 137–173.

Denes, P. B. (1963). On the statistics of spoken English. *The Journal of the Acoustical Society of America*, *35*(6), 892-904.

Dilley, L. C., & Pitt, M. A. (2007). A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. *The Journal of the Acoustical Society of America*, *122*(4), 2340-2353.

Dilley, L. C., Millett, A. L., McAuley, J. D., & Bergeson, T. R. (2014). Phonetic variation in consonants in infant-directed and adult-directed speech: the case of regressive place assimilation in word-final alveolar stops. *Journal of Child Language*, 41(1), 155.

Ernestus, M., Lahey, M., Verhees, F., & Baayen, R. H. (2006). Lexical frequency and voice assimilation. *The Journal of the Acoustical Society of America*, *120*(2), 1040-1051.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review*, *120*(4), 751.

Ferguson, C. A. (1964). Baby talk in six languages. American anthropologist, 66(6_PART2), 103–114.

Hallé, P. A., & de Boysson-Bardies, B. (1996). The format of representation of recognized words in infants' early receptive lexicon. *Infant Behavior and Development*, *19*(4), 463-481.

Herd, W., Jongman, A., & Sereno, J. (2010). An acoustic and perceptual analysis of/t/and/d/flaps in American English. *Journal of Phonetics*, *38*(4), 504-516.

Hohne, E. A., & Jusczyk, P. W. (1994). Two-month-old infants' sensitivity to allophonic differences. *Perception & Psychophysics*, 56(6), 613-623.

Johnson, K. (2004) Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (eds.) *Spontaneous Speech: Data and Analysis*. *Proceedings of the 1st Session of the 10th International Symposium*. Tokyo, Japan: The National International Institute for Japanese Language. pp. 29-54.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, 39(3-4), 159-207.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684-686.

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Cengage learning.

Lahey, M., & Ernestus, M. (2014). Pronunciation variation in infant-directed speech: Phonetic reduction of two highly frequent words. Language Learning and Development, 10(4), 308–327.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. Cognitive science, 37(1), 103-124.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological science*, *26*(3), 341-347.

Malécot, A., & Lloyd, P.M. (1968). The /t/−/d/ distinction in American alveolar flaps. Lingua, 19, 264-272.

Odden, D. (2005). Introducing phonology. Cambridge University Press, New York.

Patterson, D., & Connine, C. M. (2001). Variant frequency in flap production. *Phonetica*, *58*(4), 254-275.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191-196.

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive science*, *39*(8), 1824-1854.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89-95.

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, *53*(1), 181-216.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology*: Human perception and performance, 20(2), 421.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). FAVE (forced alignment and vowel extraction) program suite v1.2.2 10.5281/zenodo.22281.

Seidl, A., Cristià, A., Bernard, A., & Onishi, K. H. (2009). Allophonic and phonemic contrasts in infants' learning of sound patterns. *Language Learning and Development*, 5(3), 191-202.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.

Shockey, L., & Bond, Z. S. (1980). Phonological processes in speech addressed to children. *Phonetica*, 37(4), 267-274.

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech?. *Developmental science*, 15(5), 659-673.

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning?. *Journal of Child Language*, 40(3), 672.

Snoeren, N. D., Hallé, P. A., & Segui, J. (2006). A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics*, *34*(2), 241-268.

Tobias, J. V. (1959). Relative occurrence of phonemes in American English. *The Journal of the Acoustical Society of America*, 31(5), 631-631.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, 24(11), 2143-2152.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1), 49-63.

Zue, V. W., & Laferriere, M. (1979). Acoustic study of medial/t, d/in American English. *The Journal of the Acoustical Society of America*, *66*(4), 1039-1050.

Appendix A | Annotation landmarks

Phoneme	Variant	Acoustic Landmarks & Identification
/t/	[t]	audible alveolar closure; lack of voicing; audible release; lack of audible aspiration, and short VOT (usually <50ms)
	[ɾ]	periodic wave, or clear voicing bar in spectrogram; lack of formants; decreased amplitude compared to surrounding vowels; brief occlusion (≤50ms generally)
	[t^h]	identified primarily with perception; longer VOT than [t]; noticeable sibilance after stop release, particularly over 4kHz and around 2kHz
	[ʔ]	audible glottal closure or creaky voice; creaky voice around the closure; lack of sibilance/alveolar frication upon release
	[t̚]	audible alveolar closure; lack of voicing; no audible release
	[h]	no periodic wave; no stop; noisy signal distributed across frequencies; low amplitude; ghost formants of surrounding vowels
	[c]	like [t] but with audible palatal closure
	[tʃ]	S with stop release directly preceding it
/d/	[d]	audible alveolar closure; voicing during at least part of closure, unless in onset position; audible release; short VOT if in onset position (<50ms); lack of audible aspiration;
	[ɾ]	periodic wave, or clear voicing bar in spectrogram; lack of formants; decreased amplitude compared to surrounding vowels; brief occlusion (≤50ms generally)
	[d̚]	audible alveolar closure; voicing during closure; no audible release

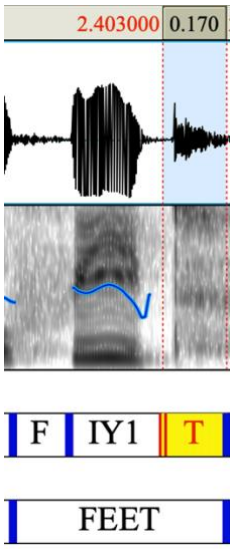
	[c]	like [t] but with audible palatal closure
	[ɟ]	like [d] but with audible palatal closure
/s/	[s]	fully continuous segment; aperiodic, high frequency noise (beginning between 5000-8000 Hz)
	[ts]	s with stop release directly preceding it
	[ʃ]	fully continuous segment; aperiodic, high frequency noise (beginning between 3000-5000 Hz)
/z/	[z]	fully continuous segment; aperiodic, high frequency noise (beginning between 5000-8000 Hz**)**
	[ʒ]	fully continuous segment; high frequency noise (beginning between 3000-5000 Hz); voicing bar, or periodic waveform
	[s]	fully continuous segment; aperiodic, high frequency noise (beginning between 5000-8000 Hz)
/n/	[n]	quieter, periodic waveform; often with antiformants; alveolar closure perceived; nasality perceived
	[n̪]	quieter, periodic waveform; often with antiformants; dental closure perceived; nasality perceived
	[ɾ̃]	periodic wave or clear voicing bar in spectrogram; decreased amplitude compared to surrounding segments; brief occlusion (≤50ms generally)
	[ɲ]	quieter, periodic waveform; often with antiformants; palatal closure perceived; nasality perceived

Appendix B | Sample spectrograms of variants

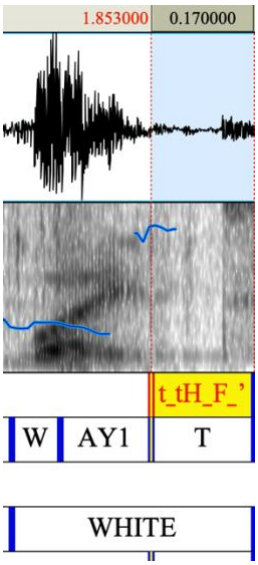
Note: because these are taken from naturalistic in-home recordings, there is background noise in each spectrogram. These were some of the best examples from our data set.

/t/

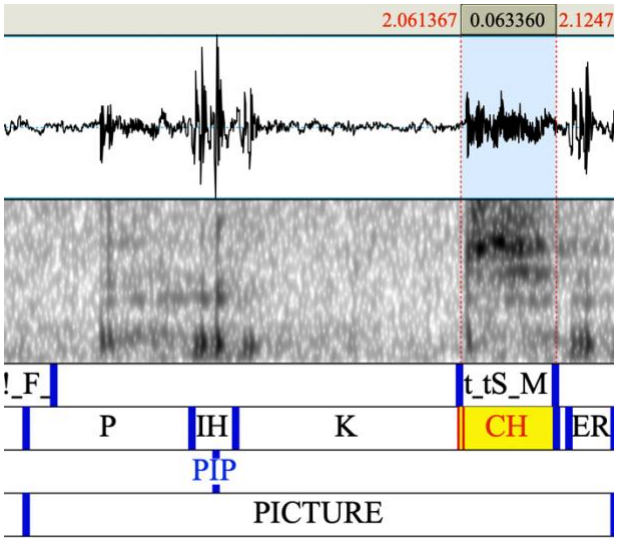
Released, unaspirated:



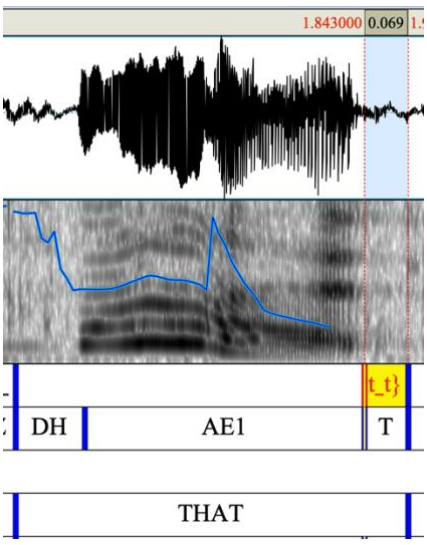
Released, aspirated:



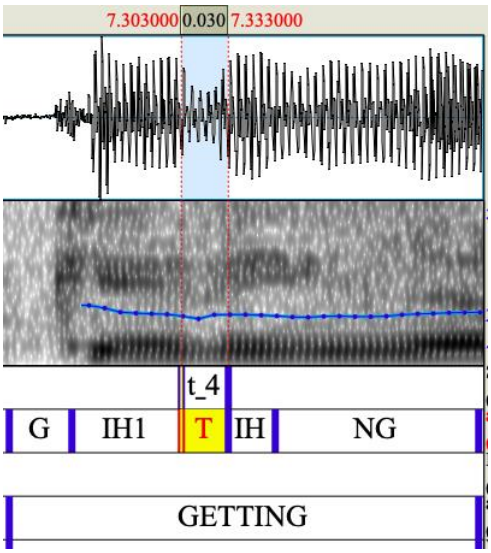
Affricated:



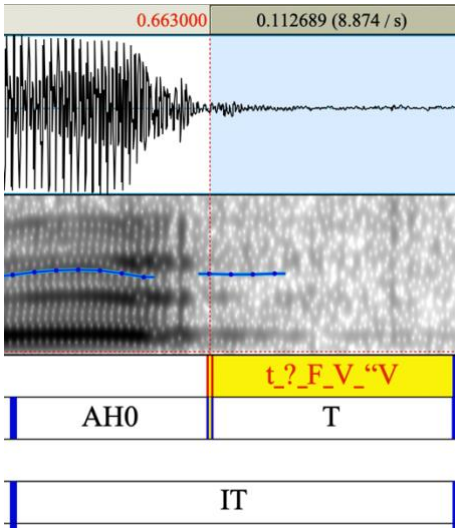
Unreleased:



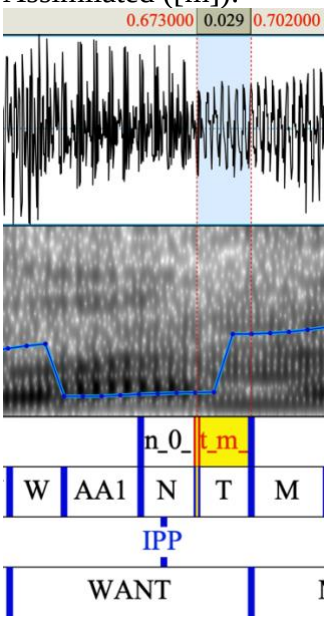
Tap:



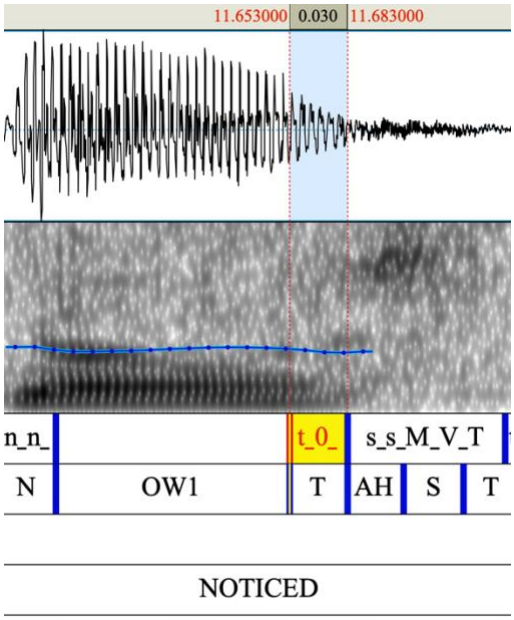
Glottal Stop:



Assimilated ([m]):

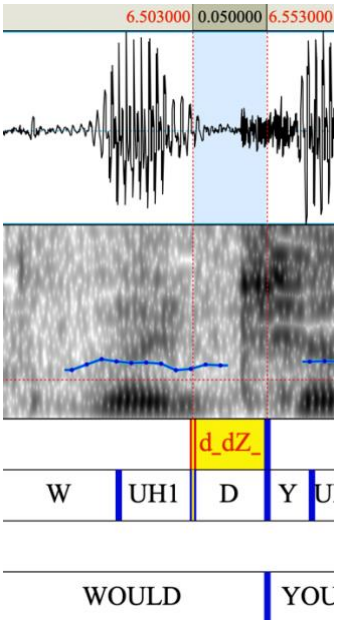


Deleted:

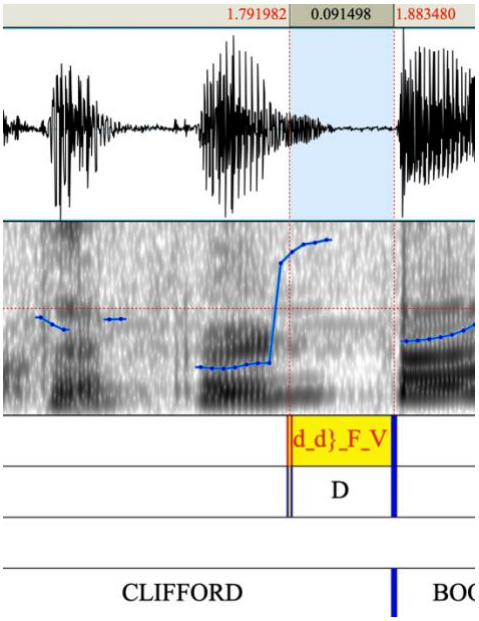


/d/

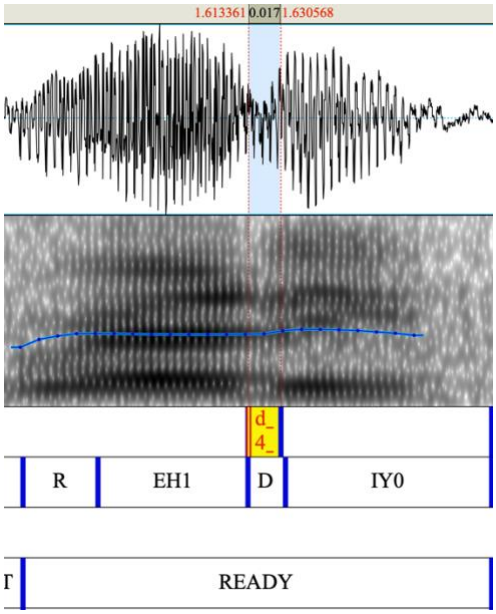
Affricated:



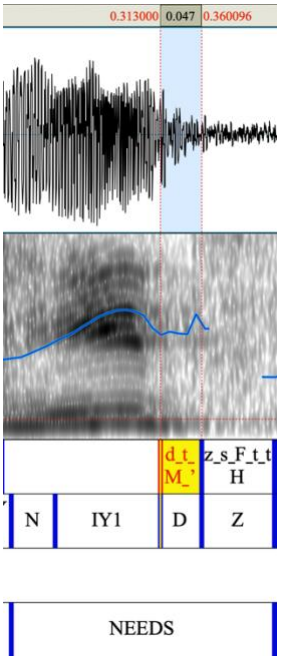
Unreleased:



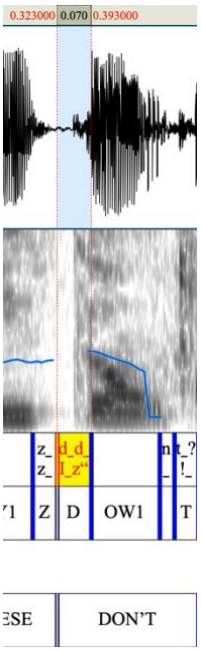
Tap:



Devoiced:

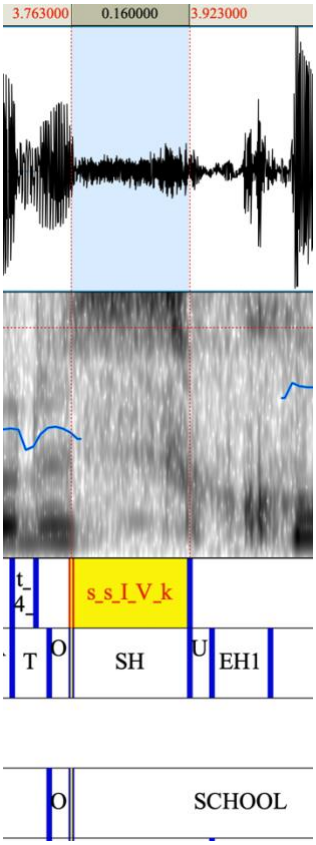


Released:

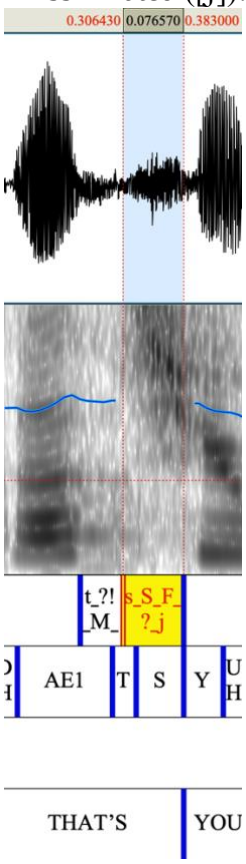


/s/

Canonical:

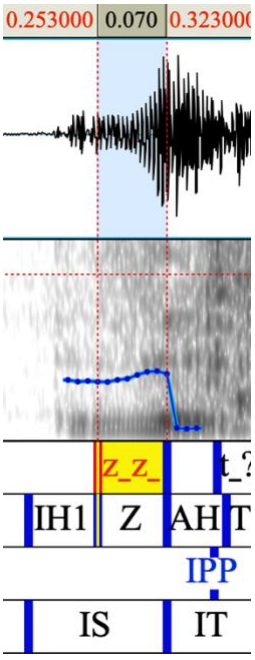


Assimilated ([ʃ]):

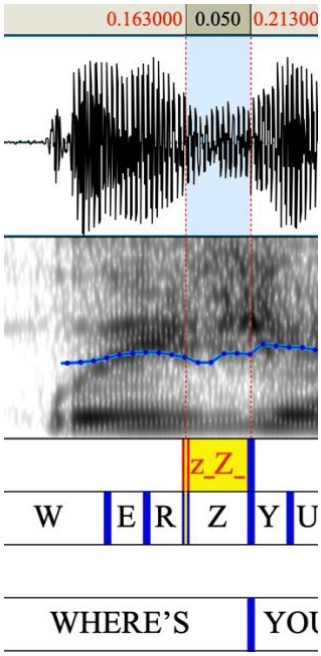


/z/

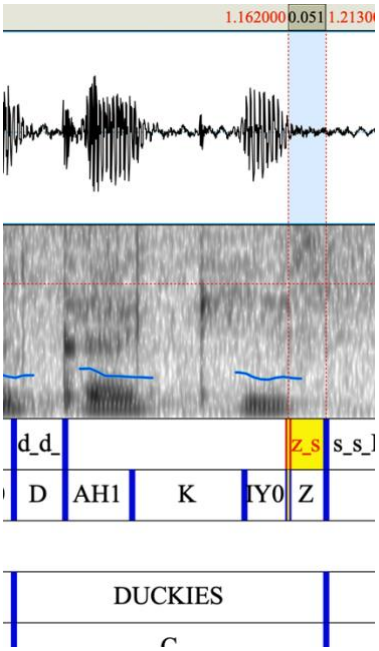
Canonical:



Assimilated ([ʒ]):

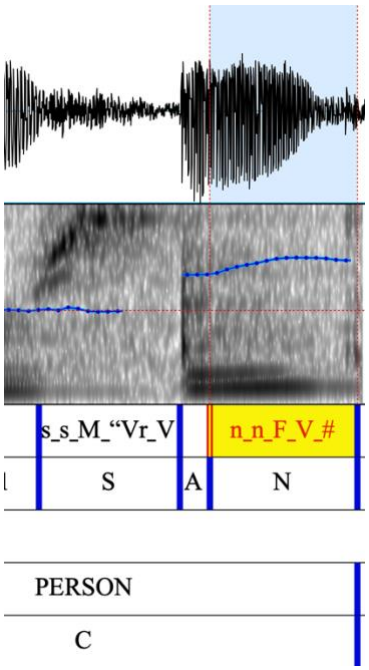


Devoiced:

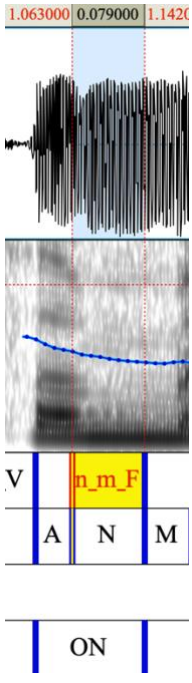


/n/

Faithful:



Assimilated ([m]):



Appendix C | Raw counts of variants

The following tables list the raw counts of the phonetic variants for each segment, overall (across all word positions) and separated by position in word.

/t/

Overall		**Initial**		**Medial**		**Final**	
Affricated	235	Affricated	113	Affricated	51	Affricated	71
Aspirated	1816	Aspirated	1382	Aspirated	159	Aspirated	275
Assimilated	38	Assimilated	1	Assimilated	1	Assimilated	36
Deleted	1630	Deleted	41	Deleted	304	Deleted	1285
Devoiced	2	Devoiced	0	Devoiced	0	Devoiced	2
Faithful	2411	Faithful	161	Faithful	1492	Faithful	759
Glottal Stop	2255	Glottal Stop	2	Glottal Stop	31	Glottal Stop	2222
Glottalized	223	Glottalized	0	Glottalized	20	Glottalized	203
Other	184	Other	29	Other	26	Other	129
Tapped	1269	Tapped	67	Tapped	612	Tapped	590
Unreleased	335	Unreleased	1	Unreleased	124	Unreleased	210
Total	*10398*	*Total*	*1797*	*Total*	*2820*	*Total*	*5782*

/d/

Overall		**Initial**		**Medial**		**Final**	
Affricated	137	Affricated	40	Affricated	5	Affricated	92
Aspirated	49	Aspirated	6	Aspirated	1	Aspirated	42
Assimilated	4	Assimilated	0	Assimilated	0	Assimilated	4
Deleted	1228	Deleted	65	Deleted	177	Deleted	986
Devoiced	182	Devoiced	84	Devoiced	28	Devoiced	70
Faithful	2675	Faithful	1526	Faithful	403	Faithful	746
Glottal Stop	5	Glottal Stop	0	Glottal Stop	0	Glottal Stop	5
Glottalized	2	Glottalized	0	Glottalized	1	Glottalized	1
Other	57	Other	16	Other	3	Other	38
Tapped	606	Tapped	102	Tapped	342	Tapped	162
Unreleased	260	Unreleased	0	Unreleased	27	Unreleased	233
Total	*5205*	*Total*	*1839*	*Total*	*987*	*Total*	*2379*

/n/

Overall		Initial		Medial		Final	
Affricated	0	Affricated	0	Affricated	0	Affricated	0
Aspirated	0	Aspirated	0	Aspirated	0	Aspirated	0
Assimilated	100	Assimilated	3	Assimilated	60	Assimilated	37
Deleted	389	Deleted	9	Deleted	252	Deleted	128
Devoiced	0	Devoiced	0	Devoiced	0	Devoiced	0
Faithful	6725	Faithful	1206	Faithful	2651	Faithful	2868
Glottal Stop	3	Glottal Stop	1	Glottal Stop	1	Glottal Stop	1
Glottalized	0	Glottalized	0	Glottalized	0	Glottalized	0
Other	27	Other	3	Other	11	Other	13
Tapped	23	Tapped	0	Tapped	10	Tapped	13
Unreleased	0	Unreleased	0	Unreleased	0	Unreleased	0
Total	*7267*	*Total*	*1222*	*Total*	*2985*	*Total*	*3060*

/s/

Overall		Initial		Medial		Final	
Affricated	20	Affricated	6	Affricated	2	Affricated	12
Aspirated	2	Aspirated	0	Aspirated	0	Aspirated	2
Assimilated	28	Assimilated	1	Assimilated	4	Assimilated	23
Deleted	41	Deleted	6	Deleted	2	Deleted	33
Devoiced	0	Devoiced	0	Devoiced	0	Devoiced	0
Faithful	6292	Faithful	2491	Faithful	750	Faithful	3051
Glottal Stop	2	Glottal Stop	0	Glottal Stop	0	Glottal Stop	2
Glottalized	0	Glottalized	0	Glottalized	0	Glottalized	0
Other	39	Other	8	Other	2	Other	29
Tapped	0	Tapped	0	Tapped	0	Tapped	0
Unreleased	1	Unreleased	1	Unreleased	0	Unreleased	0
Total	*6425*	*Total*	*2513*	*Total*	*760*	*Total*	*3152*

/z/

Overall		Initial		Medial		Final	
Affricated	11	Affricated	2	Affricated	0	Affricated	9
Aspirated	1	Aspirated	0	Aspirated	0	Aspirated	1
Assimilated	90	Assimilated	0	Assimilated	0	Assimilated	90
Deleted	36	Deleted	0	Deleted	1	Deleted	35
Devoiced	758	Devoiced	0	Devoiced	23	Devoiced	735
Faithful	2996	Faithful	53	Faithful	205	Faithful	2738
Glottal Stop	1	Glottal Stop	0	Glottal Stop	1	Glottal Stop	0
Glottalized	0	Glottalized	0	Glottalized	0	Glottalized	0
Other	31	Other	0	Other	1	Other	30
Tapped	0	Tapped	0	Tapped	0	Tapped	0
Unreleased	1	Unreleased	0	Unreleased	0	Unreleased	1
Total	*3925*	*Total*	*55*	*Total*	*231*	*Total*	*3639*

Appendix D | Phonotactic Rules

Predominantly based on *A Course in Phonetics* (Ladefoged and Johnson 2014)

/t/:

1. Alveolar stops are deleted if the preceding segment is a consonant and the following segment is a consonant
2. Alveolar plosives become taps if the preceding segment is a stressed vowel and the following vowel is unstressed
3. Alveolar stops become glottal stops if before an alveolar nasal or phrase-finally
4. Stops are unreleased if the following segment is a stop
5. Voiceless alveolar plosive is aspirated if word-initial and not preceded by /s/
6. Alveolar consonants become dental if the following segment is a dental consonant.

/d/:

1. Voiced stops are voiceless when syllable-initial, except when immediately preceded by a voiced sound
2. Stops are unreleased if the following segment is a stop
3. Alveolar stops become glottal stops if before an alveolar nasal or phrase-finally
4. Alveolar plosives become taps if the preceding segment is a stressed vowel and the following vowel is unstressed
5. Alveolar consonants become dental if the following segment is a dental consonant.