

## CHAPTER 3

### THEORIES OF PHONOTACTIC EFFECTS IN SPEECH PERCEPTION

#### **3.1. Introduction**

This chapter discusses three models of phoneme perception: the TRACE model (which puts all phonotactic effects in the lexicon), the transitional-probability model of Pitt & McQueen (1998) (which assigns them to a statistically sensitive prelexical module), and a perceptual model based on Optimality-Theoretic grammar.

#### **3.2. TRACE (McClelland & Elman 1986)**

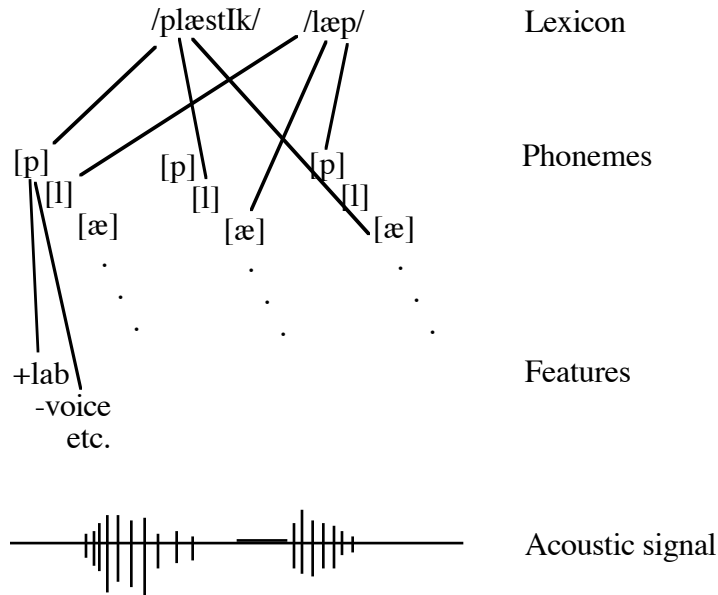
TRACE is a connectionist model of word and phoneme recognition in which the lexicon can directly influence the perception of phonemes. In TRACE, a fully or partially activated word candidate provides support for phonemic candidates which is indistinguishable from the support provided by incoming acoustic information. Phonotactic effects on speech perception are taken to arise from lack of lexical support for non-occurring phoneme sequences.

##### **3.2.1. How TRACE works**

The TRACE network is described by McClelland & Elman (1986); I will briefly summarize what they say, but for full details the reader is referred to the original paper. Each unit is a "detector" representing a hypothesis about the utterance — that it begins with a voiced sound, that it begins with a [j], that it contains the word "yard", and so forth. The detectors are organized into three layers, corresponding to features, phonemes, and words. The "activation level" of a unit is a nonnegative number which varies over time in response

to the unit's inputs. It tells how much credence the model puts in that hypothesis at the moment.

Figure 3.1. The TRACE model of McClelland and Elman (1986).



A unit receives input from all other units with which it is connected. The input which Unit A contributes to Unit B depends on Unit A's activation and another parameter, the "strength" of the A-to-B connection. All connections go both ways, so a large positive strength means A and B strongly excite each other, while a large negative strength means they strongly inhibit each other. Units on the same level inhibit each other, so that more confidence in the "plastic" hypothesis means less confidence in the "lap" hypothesis (and vice versa). Connections between levels are excitatory, so that more confidence in "plaid" means more confidence that the word starts with [p] (and vice versa). The strength of the connections is set by the experimenters; TRACE is not a learning model (McClelland & Elman 1986).

At the very bottom of the model are the acoustic feature detectors, which receive inputs not only from other units but from outside the model. Each detector is responsible

for a particular feature ([voice], [acute], etc.) at a particular time in the utterance. As the utterance unfolds moment by moment, the feature detectors register its acoustic properties and adjust their activation levels accordingly. Activation spreads upwards through the network. It also spreads back down from the word detectors to the phoneme detectors, and from them to the feature detectors. Meanwhile, the units on each level are trying to inhibit each other.

TRACE assumes that the units are open to conscious introspection: To detect X, the listener uses the X detector unit. Responses to a phoneme-monitoring task, for instance, depend on the activation levels of the phoneme units. Responses to a word-recognition task depend on the activation levels of the word units. Because activation spreads downwards and inhibition spreads sideways, a unit's activation depends not just on the acoustical configuration which it is nominally supposed to detect, but on the state of the rest of the network. Under the right circumstances, the result can be strong activation (inhibition) of the X detector despite the absence (presence) of evidence for X in the acoustic signal — a perceptual illusion.

TRACE puts phonotactic illegality in the lexicon. Legal and illegal sequences are processed differently because the legal ones receive support from lexical items containing them, while the illegal ones do not (since, by definition, they do not appear in any words) — that is, instead of punishing illegality, TRACE rewards legality. TRACE thus cannot distinguish illegal sequences from other sequences of zero frequency. Any behavioral differences between processing of zero-frequency legal sequences and illegal sequences (if they can be shown to exist) would have to be explained by something outside of the TRACE system.

### **3.2.2. Lexical effects on phoneme perception**

The lexicon can certainly influence performance on tasks that are intended to tap phoneme perception, lending credence to the TRACE approach. Evidence comes from four major paradigms:

*Phoneme detection (Foss 1969)*. Subjects listen to each stimulus and respond "yes" or "no" depending on whether it has or lacks a particular sound (usually specified as a letter). The usual dependent measure is RT for correct detections; error rates are < 10% and not useful.

*Phoneme categorization (Liberman, Harris, Hoffman & Griffith 1957)*. The stimulus is acoustically ambiguous (e.g., between [bin] and [pin]); subjects are asked which one it sounds more like. Dependent measures vary; a common one is the point (e.g., on the VOT continuum) at which both judgments are equally likely. RT is also measured, and tends to peak at the category boundary.

*Phoneme restoration (Samuel 1981ab)*. One phoneme of the stimulus (non)word has been either replaced by noise or obscured by noise, and the subject has to say which. Dependent measures are signal-detection-theoretic d-primes and betas. The effects are robust; performance is not improved by 10 000 trials of practice, nor by any but the most explicit preview cuing (Samuel 1991).

*Shadowing (Cherry 1953)*. Subject hears speech over headphones, and has to repeat it in as close to real time as possible. Various dependent measures evaluate how well mispronunciations were detected.

TRACE attributes lexical effects on phoneme perception to downward spreading of activation or inhibition from the word units to the phoneme units (McClelland & Elman 1986). Phonetic data extracted from the speech stream is only one influence on the phoneme units; it can be drowned out by powerful signals from above which bias a phoneme unit so strongly in one direction that conflicting information from below is not enough to offset it. The phoneme unit's activation level is trapped between a fixed minimum and maximum, and becomes less responsive to its inputs the closer its activation level is to

the floor or ceiling; hence, strong excitation or inhibition from above can also reduce a phoneme detector's sensitivity to acoustic features.

The three lexical factors known to influence phoneme tasks are lexicality, frequency, and uniqueness point (UP).

*Lexicality and frequency.* Effects are rather fragile for many paradigms. The only reliable one is the Ganong effect in phonetic categorization: If a stimulus is ambiguous between a word and a nonword owing to ambiguity in one phoneme, the phoneme tends to be heard so that it makes the word (Ganong 1980, Fox 1984, Connine & Clifton 1987, McQueen 1991, Pitt & Samuel 1993; not replicated by Burton, Baum, & Blumstein 1989). The same effect was observed in shadowing by Marslen-Wilson (1984): Subjects "fluently restored" mispronounced words, apparently without noticing the discrepancy (i.e., with no effect on shadowing latency). There is at least one report that a one-phoneme ambiguity between a common and a rare word tends to be resolved in favor of the common word. However, the effect can be reversed (to favor the *rarer* word) by setting up the experiment so that the less common word tends to be the right answer (Connine, Titone, & Wang 1993).

Phoneme targets *may be* detected faster in words (Rubin, Turvey, & van Gelder 1976, Cutler, Mehler, Norris, & Segui 1987; not replicated by Foss, Harwood, & Blank 1980, Frauenfelder, Segui, & Dijkstra 1990). Word initial phonemes *may be* quicker to detect in common than rare words (Morton & Long 1976, Dell & Newman 1980; not replicated by Segui & Frauenfelder 1986). Phoneme restoration *may be* stronger in words than in nonwords (Samuel 1981a, 1996; not replicated by Samuel 1987), and in common than rare words (Samuel 1981a; not replicated by Samuel 1981b).

TRACE attributes the word-superiority and shadowing-correction effects to downward spread of lexical activation. The ambiguous acoustic stimulus excites, say, [t] and [d] equally in the context *yar\_*. Since the YARD unit is somewhat activated by the context, it contributes activation to [d]; lacking stimulation from a \*YART, [t] is overtaken

by [d]. The YARD and [d] units keep exciting each other and inhibiting [t] until [t] is completely overwhelmed. A stimulus ambiguous between two nonwords, like *sirt* and *sird*, would favor no word node over any other, and would be decided on the basis of the acoustic evidence (McClellan & Elman 1986). Similar reasoning would apply if YART were a real word, but much less frequent than YARD.

The possible effects on detection are accounted for by TRACE: Top-down activation spreading causes the activation of the relevant phoneme unit to reach response criterion sooner.

*Uniqueness point* (Marslen-Wilson 1984). RT to detect a phoneme (measured from the phoneme) decreases later in real words but not nonwords (Marslen-Wilson 1984, Frauenfelder, Segui & Dijkstra 1990, Wurm & Samuel 1997); effect size varies from ~30 ms to ~300 ms. Phoneme restoration is stronger late in words with early uniqueness points than late in words with late uniqueness points (Samuel 1987). Shadows restore mispronunciations more later in the word (Marslen-Wilson & Welsh 1978). Response time to reject a nonword is a fixed amount, measured from the earliest point where the stimulus differs from all words in the dictionary (Marslen-Wilson 1984).

Strong support for the COHORT word-recognition model comes from these effects, which are hard to explain in other theories, and it is a great virtue of TRACE that the connectionist architecture is able to do that. Simulations show that an active word unit is quickly extinguished when mismatching acoustic information comes in, provided that a better-matching word unit is present (McClelland & Elman 1986). Shortly after the uniqueness point, only the matching word unit is still active, strongly inhibiting all rivals and exciting phoneme units consistent with it (which in turn inhibit phoneme units that are *inconsistent*). If mismatching phonetic information comes in after the uniqueness point, it will have a hard time changing the network's mind about the word or the phonemic code.

### **3.2.3. Phonotactic effects on phoneme perception**

In the previous section we saw that a listener's performance on a phonemic task can be influenced by their knowledge of the real words of their language. Interestingly, evidence exists that it can be influenced by their knowledge of the *possible* words of their language.

Massaro and Cohen (1983) created segments ambiguous between [r] and [l] by varying F3, and asked subjects to judge them in the contexts [t\_i], [p\_i], [v\_i], and [s\_i]. In English, only [r] is permissible after [t], only [l] is permissible after [s], both can follow [p], and neither can follow [v]. The ambiguous segments were most likely to be judged [r] in [t\_i], less likely in [p\_i], less likely still in [v\_i] and [s\_i] (as shown in their Figure 3.1). Despite the lexical confounds — [tri], [pri], and [pli] are words — the evidence of [v\_i] and [s\_i] suggests that judgments are altered by people's knowledge that [sri] can't be English: The larger number of [r] judgments after [v] than [s] cannot be due to acoustic-phonetic factors — if anything, the labial [v] should make the following F3 sound *higher* and the ambiguous segment more [l]-like — nor to lexical ones, since [sli] is not a word — which leaves phonotactics.

The TRACE model has a good explanation for how phonotactics exerts this influence. McClelland & Elman (1986) found that the ambiguous stimulus [s?i] partially activates similar words in the lexicon of their simulated word recognizer. Since the lexicon contains only phonotactically permissible words, the units for *sleep*, *sleet*, and so on become active, feeding excitation back to [l], but no countervailing *sreep* or *sreet* assists [r]. The amount of acoustic support that [l] needs to reach criterion is thus reduced in the context [s\_i] compared to a neutral context like [v\_i] in which there are no lexical items similar enough to be activated. A phonotactic effect is achieved without phonotactic rules.

#### **3.2.4. Empirical shortcomings of TRACE**

Since TRACE models many different things, it has been criticized on many different grounds. For a comprehensive review of its shortcomings as a model of phoneme

perception, see McQueen, Norris, & Cutler (1999). Most important, for our study, is that the phonotactic effect is usually larger and more robust than the lexical effect. This is very unexpected in a theory which takes phonotactic effects to be diluted lexical effects.

The evanescence of lexical effects came up in §2.2. In a lengthy study, Cutler et al. (1987) found that they could make word-superiority effects come and go by boring the listeners less or more. A varied stimulus set, containing mono- and disyllables, got the lexical effect; a monotonous one did not. They concluded that the lexical effect was not automatic, but depended on listeners' allocation of attention between the lexical and the prelexical levels of representation. For a detailed review of such results from several paradigms, see McQueen et al. (1999).

Phonotactic effects, by contrast, are robust and not affected by stimulus monotony. The original Massaro & Cohen (1983) experiment got very large effects with a monosyllabic stimulus set repeated for 1120 trials (total over two days). Pitt (1998) got several large phonotactic effects with monosyllabic stimuli. Moreton & Amano (1999) directly compared the effect of lexical status on the Japanese vowel-length boundary with that of phonotactics using the same subjects and paradigm. They found a large phonotactic effect but barely any lexical effect.

In other words, manipulations that make the lexical effect go away can still leave a phonotactic effect. This is a problem for any theory which, like TRACE, denies a distinction between lexical and phonological gaps.<sup>1</sup>

---

<sup>1</sup> McClelland and Elman (1986) report an experiment which suggests that the lexical effect and phonotactic effects can be superimposed. They compared listeners' judgments of a segment [ʔ] between [b] and [d] in the contexts *\_windle*, *\_wiffle*, and *\_wacelet*. The highest rate of "d" response was obtained in the *\_windle* context, where *dwindle* is a word; an intermediate rate was found in *\_wiffle*, where neither endpoint is a word; and the lowest rate was found in *\_wacelet*, where neither endpoint is a word but *bwacelet* is very similar to one. They interpreted this as a lexical effect superimposed on a phonotactic bias against \*[bw] – i.e., even though all the contexts were phonotactically biased, a lexical effect was still obtained.

Since all of the contexts started with *\_w*, the experiment did not demonstrate a phonotactic bias; its presence was simply assumed. As shown in Chapter 4, English listeners' bias against [bw] is weak if it exists at all. Hence, the experiment may just have measured an isolated lexical effect. It could still be true that a really strong phonotactic effect, like the bias against initial [dl], would swamp any lexical effect.

In TRACE, the phonotactic effect could be stronger because it combines the effects of many lexical items, while the lexical-superiority effect depends on a single item. However, the TRACE authors have shown that in fact the lexical-superiority effect is stronger: When the network is presented with an ambiguous phoneme between [p] and [t] in the context [\_luli], it is classified as [t], with the lexical influence of *truly* winning out over the phonotactic badness of [tl] (McClelland & Elman 1986).

### **3.3. The MERGE Transitional Probability theory (Pitt & McQueen 1998)**

Another potentially exploitable redundancy in speech is the statistical distribution of segments. Different segment sequences, such as diphones or triphones, occur with different frequencies. A model which is sensitive to these frequencies can compare the statistical plausibility of alternative parses of ambiguous speech input in order to disambiguate it. Such statistical information can also in principle be used to find word boundaries, and serve as the basis for possible-word judgments.

There is evidence from various sources that listeners are sensitive to sequence frequency. Treiman et al. (1996) found that nonwords containing high-probability sequences are rated as "more English-like" by native speakers than those containing low-probability sequences, and that, when subjects are asked to construct portmanteaus by blending two nonwords, low-frequency sequences tend to be broken up more often than high-frequency ones. Frisch et al. (2000) showed that listeners' "wordlikeness" judgments were very strongly affected by the frequency of legal phoneme sequences contained in the stimulus. Vitevich et al. (1997) found that nonwords containing frequent sequences were rated as "more English-like" and were repeated faster in a single-word shadowing task than nonwords containing rare sequences. English listeners learning an artificial language develop statistical sensitivity to the different probabilities of sound sequences in that language even if the linguistic input is an unattended background stimulus (Saffran et al. 1996, 1997; Aslin et al. 1998).

Some evidence that the sequences are encoded separately from the lexicon comes from work by Vitevich & Luce (1998, Experiment 1). They constructed lists of disyllabic English words and nonwords which varied in sequence frequency, so that some items were "high probability" and some were "low probability". When pairs of nonwords were presented for same-different judgments, listeners responded faster to high-probability pairs than low-probability pairs. When pairs of words were presented, however, the pattern was reversed: faster for low-probability, slower for high-probability. The authors' interpretation is that high-probability words and nonwords are both facilitated by the frequency of their sublexical sequences, but the high-probability words are more strongly inhibited by competition from their many lexical neighbors. Further evidence is provided by Pitt & McQueen (1998), in which an ambiguous fricative disambiguated by lexical information (the Ganong effect) did not induce compensation for coarticulation in a following ambiguous stop, while a fricative disambiguated by diphone-frequency information did.

Where TRACE seeks to explain phonotactic illegality as a gap in the lexicon, the probabilistic theories seek to explain it as a gap in the set of attested sequences: A phonotactically illegal configuration is one which has zero frequency (Pitt & McQueen 1998:349). Like the TRACE account, a probabilistic theory predicts (1) that illegal sequences are only slightly different from rare sequences, and (2) that all zero-frequency sequences (of the relevant length) are equally illegal.

In this study, we will focus on one particular implementation of a probabilistic theory, namely, that of Pitt & McQueen (1998), because it is the one which was designed for problems of ambiguous phoneme perception. The authors actually define a class of probabilistic theories, rather than a specific one; certain manipulable parameters are left unfixed. This section will try to narrow down the range of possible implementations on the basis of existing data, so that the remaining possibilities can be tested experimentally.

The rest of this section is organized as follows: §3.3.1. illustrates the functional utility of statistical knowledge. §3.3.2. describes the range of possible probabilistic theories

in the Pitt-McQueen class. §3.3.3. discusses these theories with respect to the existing data on ambiguous-phoneme perception, eliminates some of them, and defines the specific models which we will test.

### 3.3.1. Simulation: Success of statistical predictions

The functional motivation for probabilistic speech perception is clear: Sequence probabilities, even for very short sequences, greatly constrain the hypothesis space which the listener must search. To illustrate this, let us consider a model whose task is to listen to a list of isolated words drawn from the Celex English lemma database. The words occur with their Celex spoken corpus frequencies. Every so often, one of the words is truncated at a random location at least  $n$  segments into the word, and the model is asked to predict the next segment (word boundaries are counted as segments). For this task, the model has available only a table of transitional probabilities: for each string of  $n$  segments, it knows the likelihood that the  $n+1$ st will be [a], [t], etc. An example, for  $n = 2$ , is shown in Table 3.2:

Table 3.2. Probability that a given diphone will be followed by a given segment (extract from complete table).

Preceding context	Segment	Probability
.	.	.
.	.	.
.	.	.
wΛ	n	0.97
wΛ	r	0.03
wΛ	s	0.00
vZ	)	0.99
vZ	d	0.01

væ	ŋ	0.00
væ	g	0.00
væ	k	0.07
væ	l	0.62
væ	m	0.01
væ	n	0.25
.	.	.
.	.	.
.	.	.

---

Note: "(" and ")" mark word boundaries.

To predict which segment will follow a given context, the model's best strategy is to always guess the segment that is most frequent after that context, since that maximizes its chance of guessing right.

As an illustration, a simulation of this hypothetical experiment was run in which words were randomly chosen from the Celex wordforms database (EPW.CD) according to their frequency (combined written and spoken, which is Field 3 of EPW.CD). Initial and final word-boundary markers were added to each word's segmental representation. From each representation, a substring of length  $n + 1$  was chosen at random (if the word was long enough), and the model was asked to guess the last segment on the basis of its likelihood given the first  $n$ . The model was credited with a correct guess if the final segment of the substring was the best guess according to the guessing strategy (i.e., if the actual last segment was the likeliest). The simulation<sup>2</sup> was run for approximately 100,000 trials for each of  $n = 1, 2,$  and  $3$ . It did rather well:

Table 3.3. Results of the simulation: Success rate as a function of context size.

<sup>2</sup> The script `simulated_guess` is included in the appendix.

Size of preceding context ( <i>n</i> )	Model's success rate
1	0.378
2	0.630
3	0.783

That is, knowing only the last two segments, the model will predict the next one correctly nearly two-thirds of the time – with zero acoustic information and zero lexical information. This is only a thought experiment, but it is close enough to both the lab and real life to show that even a small amount of probability knowledge can be used to very great advantage.

### 3.3.2. Probabilistic theories of speech perception

The probabilistic theory of Pitt & McQueen (1998) is, essentially, that prelexical mechanisms are sensitive to sequence frequencies (in the equivalent form of transitional probabilities), and that, when acoustic evidence is inconclusive, perception favors the more likely option. This prelexical probabilistic module, along with the Shortlist model of word recognition (Norris et al. 2000), forms part of the MERGE model, a theory of phoneme-processing tasks in which the output of prelexical phonemic processing, along with lexical information, is used in making phoneme-based decisions (Norris, McQueen, & Cutler (in press)). Probabilistic effects, in this model, occur very early, and are separate from lexical effects. The relative contribution of each to phoneme responses is determined by attentional weighting, which in turn is determined by task variables.

There are several different ways to make a theory of sequence-probability influence on phoneme perception. The major adjustable parameters include: (1) location and size of the context, (2) the database from which the probabilities are computed, and (3) the guessing

strategy. This section describes the possible parameter settings, and the theories resulting from them.

### 3.3.2.1. Context

In a typical perception experiment, the listener is confronted with an acoustically ambiguous segment "?", which could be either  $x$  or  $y$ , in a context  $A\_B$ . How can statistical knowledge about the frequencies of  $AxB$  and  $AyB$  be used to disambiguate it?

One way is to directly compare the likelihood of  $x$  and  $y$  in the context. The decision depends on the conditional probabilities  $P(x | A\_B)$  and  $P(y | A\_B)$ , where

$$(3.4) \quad P(x|A\_B) = \frac{F(AxB)}{F(A\_B)}$$

$$P(y|A\_B) = \frac{F(AyB)}{F(A\_B)}$$

Since  $F(\text{string})$  is the frequency with which that string occurs in the database (the listener's experience), all the model has to do in order to make its decision is to compare  $F(AxB)$  with  $F(AyB)$ . It consults its table<sup>3</sup> of  $(2n+1)$ -phone frequencies, where  $n$  is the length of  $A$  and  $B$ ,<sup>4</sup> retrieves the two relevant frequencies, and hands them over to the decision rule. This kind of context I will call *surrounding context of order n*.

A second possibility is to treat the left and right contexts separately. The decision depends on the conditional probabilities  $P(x | A\_)$ ,  $P(y | A\_)$ ,  $P(x | \_B)$ , and  $P(y | \_B)$ , which reduce to the frequencies  $F(Ax)$ ,  $F(Ay)$ ,  $F(Bx)$ , and  $F(bY)$ . This I will call *independent neighboring context of order n* (again assuming simplistically that  $A$  and  $B$  have equal lengths). A table of  $(n+1)$ -phone frequencies is consulted.

---

<sup>3</sup> I say "table", but that is only one notational variant. They can also be viewed as sublexical-sequence detector units whose resting activation or excitability depends on frequency. A proposal along these lines is Luce & Vitevich (\*\*\*)

<sup>4</sup> For theoretical simplicity's sake I assume symmetry;  $A$  and  $B$  have the same length. This might be wrong.

The predictive difference between these two context types is that surrounding context (SC) can take advantage of statistical dependencies between *A* and *B*, while independent neighboring context (INC) cannot.

For example, English, like most languages, requires sonority to rise in syllable onsets but not in codas. As a result, it lacks sequences like [tdt], [pdp], [fvp], etc., since no matter what precedes or follows such a sequence, there is no legitimate syllabification – the consonant in the middle is higher in sonority than either of its neighbors, but not high enough that it can serve as a syllable nucleus itself. A model using SC of order 1 will note the gaps in its table of 3-phone frequencies. A model using INC of order 1, though, will miss these gaps, since each of the sub-sequences [td], [dt], etc. does in fact occur. Given a segment ambiguous between [l] and [n] in the context [m\_z], the SC-1 model will favor [l], since [mlz] is attested (e.g., in *camels*), while [mnz] isn't (at least, not for speakers who lack syllabic [n] in *lemons*). The INC-1 model will note only the low nonzero frequency of [mn] (e.g., *damnation, amnesia*), and the high frequencies of [mz], [ml] and [lz], and treat [n] as no worse in [m\_z] than in [m\_eI]. Which of these models is closer to what people do is of course a question for the laboratory.

Another difference between SC and INC, of no consequence predictively but significant conceptually, is the size of the *n*-phone tables. As the size of the context increases, so does the number of phoneme sequences whose frequencies the prelexical module has to keep track of. Their number quickly approaches the size of the lexicon:<sup>5</sup>

Table 3.5. Attested English phoneme sequences of lengths 2, 3, and 4.

Set	Size
Celex wordforms, ≥ 1 per mio. wds:	
length-2 sequences	1,395
length-3 sequences	11,961

<sup>5</sup> The single word *bat* [bæt], for example, contributes four 2-phones: #b, bæ, æt, and t#.

length-4 sequences

35,732

---

Celex lemmas (EPL.CD),  $\geq 0$  per mio. wds

52,447

---

Note: Initial and final word boundaries were counted as phonemes. The sizes can be reduced somewhat by not counting them.

The MERGE TP models are intended to contrast with TRACE by keeping the lexicon out of the early stages of speech perception. As the  $n$ -phone tables grow, this difference becomes blurred: the tables incorporate not only the entire lexicon of length  $n$  or less, but fragments of many larger words, and the two theories come to make more and more similar predictions. These considerations argue for the INC theories over the SC theories, since the former use shorter  $n$ -phones to describe the same-sized context.

Pitt & McQueen (1998)'s experiments use a preceding context of length 1, but the authors discuss evidence that a preceding context of length 3 may be needed. Since all of the stimuli they discuss had the same following context (silence), they did not need to go into their claims about following context. I will assume that they are considering one of three theories of context: SC-1, INC-1, or INC-3. (See discussion below, §3.3.3.)

### **3.3.2.2. Database**

What corpus are the  $n$ -phone tables based on? This is both a theoretical and a practical problem. There are two principal options.

One possibility is that the  $n$ -phone frequencies are computed from the stored items in the lexicon. Each word contributes its  $n$ -phones, weighted according to the word's frequency. The lexicon does not directly participate in speech perception, but contributes off-line by updating the  $n$ -phone tables which the early perceptual mechanisms can consult. Such theories use a *lexical database*.

A second possibility is that the  $n$ -phone frequencies are computed directly from the incoming speech stream – computed by the same mechanisms that later consult them, without any participation from the lexicon. This is more in keeping with the spirit of the Pitt & McQueen (1998) model, a strictly bottom-up theory which aims to block the lexicon from interfering with the early stages of perception. Such theories use a *corpus database*.

The empirical difference between the two is that the lexical database respects morphological word boundaries, while the corpus database does not. The reason is that morphological word boundaries are represented explicitly in the lexicon, but not in a segmental analysis of the speech stream<sup>6</sup>. The effect on  $n$ -phone statistics can be substantial. For instance, geminates are very rare in the English lexicon but occur freely in running speech (*That trite talk keeps Sid dozing*).

The practical difference is that on-line dictionaries make the lexical-database statistics easy to compute, while a lack of accessible on-line phonetic corpora makes the corpus-database statistics hard to compute. The standard practice in the field is therefore to use an on-line dictionary and tacitly assume that the difference is negligible until proven otherwise. Since I can't tell what the corpus-database statistics predict, I will have to ignore that theory and focus on the lexical-database theory.

Most of the frequency counts which Pitt & McQueen relied on were computed for American English pronunciations, with frequencies apparently reckoned from the million-word written American English corpus of Kucera & Francis (1967). Celex's 18-million-word corpus is much larger, separates written from spoken English, and distinguishes

---

<sup>6</sup> Morphological word boundaries have many indirect *correlates* in the surface-level phonetic analysis of the speech stream. Since prosodic boundaries tend to be aligned with them, they often correlate with fortition (Fougeron & Keating 1997). As "prominent" positions, they tend to support phonological contrasts not available elsewhere (Beckman 1998, Smith 1999), and to undergo prominence-enhancing phonological processes (Smith forthcoming). None of these correlates, however, allows morphological word boundaries to be unambiguously located in the speech stream pre-lexically. If the  $n$ -phone tables are compiled prelexically, no character corresponding to a morphological word boundary will be present in them.

(The same problem was faced by post-Bloomfieldian structuralist linguistic theory, which demanded that grammatical analysis proceed from lower to higher levels. Harris (1951) suggested a statistical solution: Morph boundaries occur where the unpredictability of the next phoneme reaches a peak. This solution, like the TP theory, makes indirect use of lexical and grammatical information. See Newmeyer (1986:7-9) for a discussion.)

different inflected forms of the same word, but its British pronunciations are a drawback when working with American English speakers. The result is some uncertainty about what the TPs really are, and hence about what a TP-based theory would actually predict. For instance, if one wants to reckon the probability that the segment following a given vowel will be [s] or [ʃ] – a crucial case in the study of Pitt and McQueen (1998) – one can get three different estimates for each:

Table 3.6. Transitional probabilities for the stimuli of Pitt & McQueen (1998),  $n = 1$ .

Transition	Probability		
	Pitt & McQueen (1998, Table 2): written Am. Eng.	Celex: written and spoken Br. Eng.	Francis-Kucera: written Am. Eng. <sup>7</sup>
Pr ( [us]   [u_] )	0.019	0.021	0.008
Pr ( [ʊf]   [u_] )	0.010	0.009	0.009
Pr ( [ʊs]   [ʊ_] )	0.004	0.002	0.001
Pr ( [ʊf]   [ʊ_] )	0.004	0.013	0.005
Pr ( [ɹs]   [ɹ_] )	0.058	0.115	0.163
Pr ( [ɹf]   [ɹ_] )	0.007	0.009	0.007
Pr ( [eɪs]   [eɪ_] )	0.064	0.026	0.069
Pr ( [eɪf]   [eɪ_] )	0.139	0.043	0.133
Pr ( [is]   [i_] )	0.021	0.017	0.023
Pr ( [if]   [i_] )	0.002	0.001	0.001
Pr ( [ip]   [i_] )	0.020	0.021	0.018
Pr ( [it]   [i_] )	0.025	0.026	0.025
Pr ( [eɪp]   [eɪ_] )	0.015	0.008	0.017
Pr ( [eɪt]   [eɪ_] )	0.151	0.054	0.123

<sup>7</sup> Computed from the same database used by Pitt & McQueen, but apparently using a somewhat different counting method.

Agreement is tolerably good if we stick to a one-segment context (i.e., a table of length-2 sequences). The absolute magnitudes may differ by a factor of three, but the different methods generally agree as to which segment each context favors.

For safety's sake I will give frequencies using both an American English frequency dictionary similar to Pitt & McQueen's and the Celex corpus. Details of how these frequencies are computed will be found in the appendix to this chapter.

To compare the conditional probability  $P(x | A\_B)$  with  $P(y | A\_B)$  – that is, the relative chances of finding  $x$  or  $y$  in a given environment – we need only compare the frequency of  $AxB$  with that of  $AyB$ , since  $P(x | A\_B) = (\text{frequency of } AxB)/(\text{frequency of the } A\_B \text{ environment})$  and  $P(y | A\_B) = (\text{frequency of } AyB)/(\text{frequency of the } A\_B \text{ environment})$ . I will therefore report only the  $AxB$  and  $AyB$  frequency counts.

### **3.3.2.3. Decision rule**

Once the statistical information has been used to estimate the probability that a particular segment in the context  $A\_B$  is  $x$  or  $y$ , the model then has to choose one of the two. How?

The general form of the decision rule must specify the probability that the model guesses  $x$  rather than  $y$  given a stimulus  $A?B$ . The decision rule has to take into account at least two things: 1. the acoustic composition of the ambiguous segment (how close it is to  $x$  or  $y$ ), and 2. the TP statistics.

In everyday life, listeners are constantly confronted with productions of  $x$  and  $y$ . Some are clear; some are garbled in various ways. The listener may at first parse a given production "wrongly" (i.e., not as the speaker intended it), but usually the correct interpretation becomes clear shortly as the listener recognizes the speaker's intended message. The listener therefore has the feedback needed to optimize the decision rule by

adjusting its parameters. We will suppose that they do this, with the goal of maximizing their likelihood of correctly restoring the intended stimulus.

Each  $AxB$  or  $AyB$  stimulus puts the listener in a particular internal state. *Which* internal state will depend not only on the speaker's intent, but on the garbling and on the perceptual noise added by the listener's auditory system. Under the TP hypothesis, the listener's response is determined by their perceptual state and by the distributional statistics of their language. For the sake of illustration, let's assume the SC-1 statistics.

Suppose the listener, having heard a particular stimulus  $A?B$  (intended by the speaker as  $AxB$  or  $AyB$ ), is now in State  $Z$ , a state which can lead to a response of "x" or "y". The likelihood of correctly guessing the intended message is

(3.7)

$$P_c(Z) = P(\text{spkr said } AxB|Z) \cdot P(\text{guess "x"}|Z) + P(\text{spkr said } AyB|Z) \cdot P(\text{guess "y"}|Z)$$

By Bayes's Theorem,

(3.8)

$$P(\text{spkr said } AxB|Z) = \frac{P(Z|\text{spkr said } AxB) \cdot P(\text{spkr said } AxB)}{P(Z|\text{spkr said } AxB) \cdot P(\text{spkr said } AxB) + P(Z|\text{spkr said } AyB) \cdot P(\text{spkr said } AyB)}$$

$$P(\text{spkr said } AyB|Z) = \frac{P(Z|\text{spkr said } AyB) \cdot P(\text{spkr said } AyB)}{P(Z|\text{spkr said } AxB) \cdot P(\text{spkr said } AxB) + P(Z|\text{spkr said } AyB) \cdot P(\text{spkr said } AyB)}$$

Letting  $r_x = P(Z | \text{speaker said } AxB)$ ,  $p_x = P(\text{speaker said } AxB)$ ,  $q_x = P(\text{guess "x" } | Z)$ , and similarly for y, we get

(3.9)

$$\begin{aligned} P_c(Z) &= \frac{r_x p_x}{r_x p_x + r_y p_y} q_x + \frac{r_y p_y}{r_x p_x + r_y p_y} q_y \\ &= \frac{r_x p_x}{r_x p_x + r_y p_y} q_x + \frac{r_y p_y}{r_x p_x + r_y p_y} (1 - q_x) \end{aligned}$$

What choice of  $q_x$ , our only free parameter, maximizes our chance of guessing correctly? Clearly<sup>8</sup>, either  $q_x = 0$  or  $q_x = 1$ . Since  $Z$  was arbitrarily chosen, it is true in general that from any given internal state, the optimal choice is either always "x" or always "y". The choice depends on the  $r$  and  $p$  parameters: if  $r_x p_x > r_y p_y$ , then "x" is the best guess; if the reverse, then "y".

In the language of Signal Detection Theory (Green & Swets 1966, Macmillan & Creelman 1991), Choice Theory (Luce 1963), or Generalized Recognition Theory (Ashby & Maddox 1994), an acoustic stimulus evokes an internal representation as a point in a perceptual space. ("State  $Z$ " is one such point.) Following the reasoning described above, the space is partitioned into regions, and all points in the same region lead to the same response. To get optimal performance, each region must contain only points where  $r_x p_x \geq r_y p_y$ , or only points where  $r_x p_x \leq r_y p_y$ , so the boundaries must be drawn so that  $r_x p_x = r_y p_y$  for points  $Z$  on the boundary (Macmillan & Creelman 1991:Ch. 1); i.e., so that  $r_x/r_y = p_y/p_x$ , or

(3.10)

$$\frac{P(Z|\text{spkr said } AyB)}{P(Z|\text{spkr said } AxB)} = \frac{P(\text{spkr said } AxB)}{P(\text{spkr said } AyB)}$$

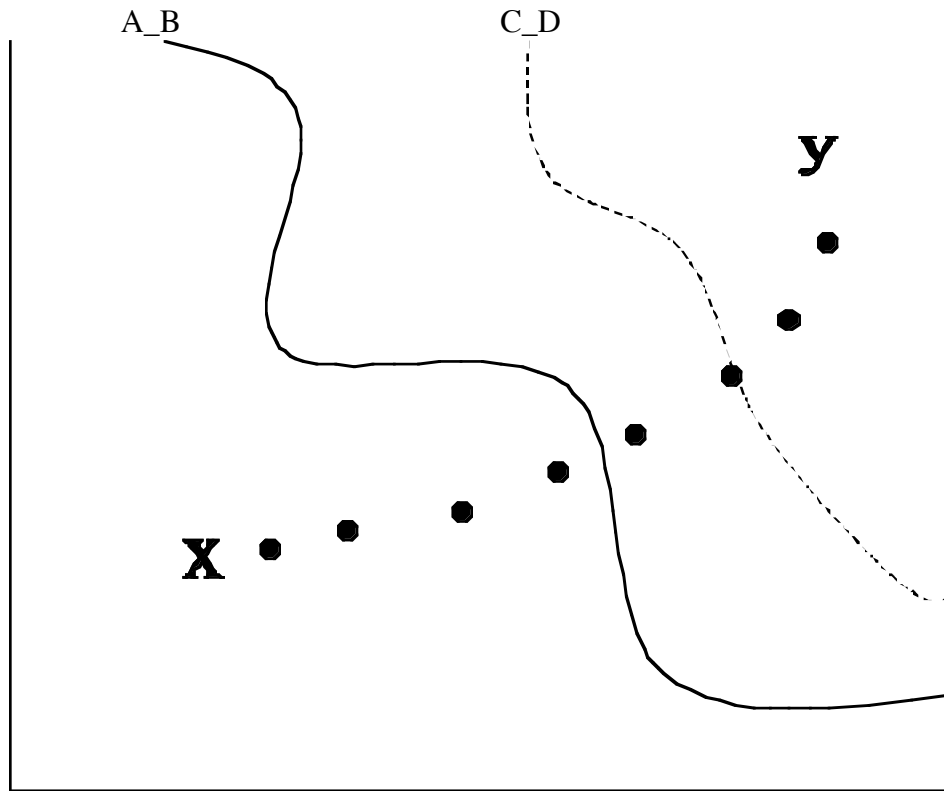
---

<sup>8</sup>  $P_c(Z)$  is linear in  $q_x$ , so its maximum must be at the smallest or largest possible value of  $q_x$ , i.e., 0 or 1.

If the a-priori probability ratio (the right-hand side) changes, as when  $A_B$  is replaced by a different phonological environment  $C_D$  in which  $y$  is less likely, then the boundary must move in order to keep the likelihood ratio (the left-hand side) equal to it.

The consequences for a typical perceptual experiment are illustrated in (10). The plane is perceptual space, with  $x$  and  $y$  the idealized perceptual representations of the endpoint stimuli – "idealized", because in fact perceptual noise causes each presentation of a stimulus to evoke a slightly different percept. The irregular line of dots shows the idealized locations of the intermediate stimuli. Following the optimal response strategy, listeners respond "x" when the percept is on one side of the boundary and "y" when it is on the other; hence, what will be observed in the experiment is that the responses cross over from mostly "x" to mostly "y" where the line of stimuli intersects the boundary. The difference in boundary location between the  $A_B$  and  $C_D$  contexts causes a corresponding shift in the location of the "x"/"y" crossover point.

Figure 3.11. Boundary shift in perceptual space.



The listener thus optimizes performance by changing their willingness to respond "x" in accordance with the ratio of the probabilities of  $x$  and  $y$  in the context. This leads to an important conclusion: The effect of context on the location of the "x"/"y" response boundary depends on the ratio of the probabilities of  $x$  and  $y$  in that context, and not on their difference.

For example: Suppose  $AxB$ ,  $AyB$ , and  $CxD$  all occur 1,000 times per million words, while  $CyD$  occurs 901 times per million words. The  $x/y$  ratio in  $A_B$  is 1, while that in  $CyD$  is about 1.1. If a shift in the "x"/"y" boundary between  $A_B$  and  $C_D$  is found experimentally, we would expect an even larger shift between  $A'_B$  and  $C'_D$ , where  $A'xB'$ ,  $A'yB'$ , and  $C'xD'$  occur 100 times per million words and  $C'yD'$  occurs 1 time per million words (giving  $x/y$  ratios of 1 in  $A'_B$  and 100 in  $C'_D$ ). Though the frequency differences are the same (99 per million in both cases), it is the ratios that matter.

It is very unlikely that listeners follow the optimal strategy to the letter, which would entail disregarding acoustic evidence of an event of zero probability. Under the optimal strategy, a sequence of probability 0 is infinitely unlikelier than a sequence of positive probability; hence, phonotactically illegal stimuli should always be as legal. There is probably a limit to how far the criterion can be shifted, so that larger and larger a-priori probability ratios only increase the bias up to a point. Very infrequent sequences are therefore expected to behave similarly to absolutely non-occurring ones.

### 3.3.3. Statistical context effects on phoneme perception

The MERGE TP theory is intended to explain an interaction between lexical and phonetic effects in phoneme perception. Mann & Repp (1981) showed that a segment ambiguous between [t] and [k] tends to be heard as [t] after [ʃ] and as [k] after [s]. The effect evidently arises at an early (low) level of processing, either because the perceptual system is compensating for expected coarticulatory effects, or because the low-frequency [ʃ] makes the next segment sound higher by contrast while the high-frequency [s] makes it sound lower (Kluender & Lotto 1994). Elman & McClelland (1988) used neither [ʃ] nor [s], but a segment [ʔ] acoustically in between them. When [ʔ] followed *Christma\_*, *ridiculou\_*, or *copiou\_*, it acted like [s] in its effect on perception of a following [t]-[k] (*tapes-capes*) continuum. When [ʔ] followed *fooli\_*, *Spani\_*, or *Engli\_*, it acted like [ʃ]. They concluded that lexical activation was spreading down to the phoneme level to favor [ʃ] or [s], as the case might be, which then had its ordinary phonetic effect on the following segment.

Pitt & McQueen (1998) argued that early phoneme processing was immune from lexical effects, and that the Elman & McClelland results could be accounted for if low-level (prelexical) phonetic processes were sensitive to segment-to-segment TPs: the ambiguous [ʔ] was behaving like [ʃ] or [s] depending on which was more likely to follow the preceding

segmental context. When [ʃ] was more likely, [ʔ] produced more [t] responses to the following [t]-[k] continuum; when [s] was more likely, [ʔ] produced more [k] responses.

Where Elman & McClelland had asked only for judgments of the [t]-[k] continuum, Pitt & McQueen (1998, Experiment 1) also asked for judgments of [ʔ]. The [ʔ] was presented in two pairs of biasing contexts. One pair, [dʒu\_] and [bʊ\_], were lexically biased towards [s] (*juice*) and [ʃ] (*bush*), but the TP from the vowel to [s] and to [ʃ] was the same for both. The other pair, [dɪ\_] and [neɪ\_], were lexically unbiased (since [dɪs], [dɪʃ], [neɪs], and [neɪʃ] are all nonwords), but differed in the TPs from the vowel to the following fricative. The relevant statistics are shown in Table 3.4, repeated here for convenience:

Table 3.12. Transitional probabilities for the stimuli of Pitt & McQueen (1998),  $n=1$ .

Transition	Probability		
	Pitt & McQueen (1998, Table 2): written Am. Eng.	Celex: written and spoken Br. Eng.	Francis-Kucera: written Am. Eng. <sup>9</sup>
Pr ( [us]   [u_] )	0.019	0.021	0.008
Pr ( [uʃ]   [u_] )	0.010	0.009	0.009
Pr ( [ʊs]   [ʊ_] )	0.004	0.002	0.001
Pr ( [ʊʃ]   [ʊ_] )	0.004	0.013	0.005
Pr ( [ɪs]   [ɪ_] )	0.058	0.115	0.163
Pr ( [ɪʃ]   [ɪ_] )	0.007	0.009	0.007
Pr ( [eɪs]   [eɪ_] )	0.064	0.026	0.069

<sup>9</sup> Computed from the same database used by Pitt & McQueen, but apparently using a somewhat different counting method.

Transition	Probability		
	Pitt & McQueen (1998, Table 2): written Am. Eng.	Celex: written and spoken Br. Eng.	Francis-Kucera: written Am. Eng. <sup>9</sup>
Pr ( [eɪʃ]   [eɪ_] )	0.139	0.043	0.133
Pr ( [is]   [i_] )	0.021	0.017	0.023
Pr ( [iʃ]   [i_] )	0.002	0.001	0.001
Pr ( [ip]   [i_] )	0.020	0.021	0.018
Pr ( [it]   [i_] )	0.025	0.026	0.025
Pr ( [eɪp]   [eɪ_] )	0.015	0.008	0.017
Pr ( [eɪt]   [eɪ_] )	0.151	0.054	0.123

Pitt and McQueen found a lexical and a TP influence on [ʔ] report, but only a TP influence on [t]-[k] report, suggesting that TPs were influencing early phonetic processing, with the lexical effect only emerging at a later stage.<sup>10</sup> At an early, prelexical stage of phonetic processing, the ambiguous fricative would be disambiguated using TPs, and, having been classified as [s] or [ʃ], would so affect the perception of the following [t] or [k]. Later on, after lexical access, the lexicon could affect listeners' report of the fricative, but not of the stop.

Does the experimental evidence from these two studies rule out any of the possible TP-based theories described in the last section? Are the left and right contexts independent

<sup>10</sup> The latter half of this squares with the findings of Fox (1984), who reported that lexical influences on ambiguous-phoneme perception turn up only among the responses with long RTs. [\*\*\* Discuss Pitt & Samuel 1993 on wordness and RT to phoneme monitoring]

(INC model), or are they treated as a single unit (SC model)? Since following context was not varied in these studies (it was always a [t]-[k] or [d]-[g] continuum before [ɛɪ]), they do not distinguish INC from SC. However, they do throw some light on the question of how much preceding context has an effect.

The most conservative hypothesis, which is used by Pitt and McQueen (1998) through most of their paper, is that only the immediately preceding segment matters: the decision between [s] and [ʃ] after [WXYZ\_] depends only on the frequencies of [Zs] and [Zʃ].

McClelland and Elman had considered this account of their results:

One might have proposed that simple phoneme-to-phoneme sequential constraints are such that they would lead subjects to predict that the final phoneme in "Spanish" was an /ʃ/ but the final phoneme in "ridiculous" was an /s/, quite apart from specific lexical factors; it may be that [nI\_] is more often completed with [ʃ], while [I?\_] <sup>11</sup> is more often completed with [s] (1988:158).

They dismiss it in view of the results of their Experiment 3, in which a large effect was found when the contexts were *fooY?* and *ridicuY?*, where "Y" was a CV sequence intermediate between [ɪ] and [ə]:

However, in the syllable-replaced condition, the context in the replaced items is actually the same for three phonemes before the final fricative; the vowel in "foo\_" and the last vowel in "ridicu\_" are the same vowel, though they may have slightly different acoustic realizations due to coarticulation, and the next two sounds in the two contexts are both acoustically and phonetically identical in the syllable-replaced stimuli. Thus, any differential prediction of the identity of the final fricative would have to be based on "f\_" vs. "ridic\_" and thus would seem to be attributable to knowledge that is specific to the particular lexical items involved (1988:158-159).

Context fully three segments away is affecting the ambiguous fricative so strongly that it in turn affects perception of the following stop. The TRACE authors argue, in effect,

---

<sup>11</sup> Sic; an apparent typo for [ə].

that expanding the TP context to be that big makes the TP theory practically lexical, by including whole words in table of sequence frequencies (e.g., all words in the lexicon which are four segments long or shorter). At any rate, a single segment of preceding context is not enough.

Pitt and McQueen reply that the last vowels in "foo\_" and "ridicu\_" are not in fact identical; the one they take to be [u] and the other to be [ʊ]. If we consider the frequencies of the four sequences [ulɪs], [ulɪʃ], [ʊlɪs], and [ʊlɪʃ], then the TPs favor [ɪs] after [ul\_] and [əʃ] after [ʊl\_].<sup>12</sup>

/ulɪʃ/ occurs in words like coolish, foolish, and ghoulishness. Celex shows that this string occurs about 26 times per million words. /ulɪs/ and /uləs/ occur less than once per million words, and /uləʃ/ does not occur at all. So /ʃ/ is much more likely given /ulV?/. The opposite bias operates after /U/. The string /Uləs/ is quite common, in words like incredulously, ridiculous, and stimulus. The CELEX estimate is 86 times per million words. /Ulɪs/ also occurs (in words like oculist and somnambulist, 9 times per million), but /ʊlɪʃ/ and /ʊlɪʃ/ never occur. So /s/ is much more likely after /UlV?/. (Pitt & McQueen 1998:365)

This counterproposal does not necessarily require the listener to keep track of 4-phones, of which there are at least 35,732 (see Table 3.1). Perhaps what has happened in McClelland and Elman's experiment is a statistical chain reaction. Suppose the listener maintains a 3-phone table (at least 11,961 entries). When the ambiguous [ɪ]/[ə] vowel is encountered after [ul\_] or [ʊl\_], it is disambiguated using 3-phones. The restored [lɪ\_] or [lə\_] context is then used to decide, statistically, between [s] and [ʃ].

The statistics of English permit this. Table (3.5) shows the relevant Celex counts for the [ɪ]/[ə] decision, which favor [ɪ] after [ul\_] and [ə] after [ʊl\_]. (The [ɪ] counts in

<sup>12</sup> *Ridiculous* is transcribed by the Francis-Kucera dictionary as *ridik[jʊ]lous*; Jones (1997) records both this and *ridic[jə]lus*. Some speakers may have this latter pronunciation. CELEX estimates /ələs/ to occur about 56 times per million words (combined spoken and written), and /ələʃ/ to occur not at all; /əlɪs/ occurs 350 times, and /əlɪʃ/ 25. Following Pitt & McQueen's reasoning, [s] is still more likely after /əlV?/.

Celex are too high for American English, since word-final unstressed [ɪ], as in *marry*, is pronounced [I] in Southern English dialects (Trudgill 1999). I have corrected them in the table by subtracting the number of word-final occurrences in each context.) Table (3.6) shows the counts for the [s]/[ʃ] decision, which strongly favor [s] after [lə\_], but are nearly neutral after [lɪ\_].

Table 3.13. Triphone frequencies for sequences ending in [ɪ]/[ə] in the stimuli of McClelland and Elman (1988).

3-phone	Frequency per million words, Celex EFW.CD/EPW.CD			Frequency per million words, Francis/Kucera
	Combined	Written	Spoken	
[ɪlɪ] (raw)	151	161	76	0 <sup>13</sup>
[ɪlɪ#] (word-final)	-72	-78	-24	
[ɪlɪ] (corrected)	79	83	52	
[ɪlə]	32	33	12	15
[ʊlɪ] (raw)	361	374	219	1
[ʊlɪ#] (word-final)	-264	-273	-171	
[ʊlɪ] (corrected)	97	101	48	
[ʊlə]	939	902	1342	175

<sup>13</sup> This American English dictionary does not contain the word *foolish*.

Table 3.6. Triphone frequencies for sequences ending in [s]/[ʃ] in the stimuli of McClelland and Elman (1988).

3-phone	Frequency per million words, Celex EFW.CD/EPW.CD			Frequency per million words, Francis/Kucera
	Combined	Written	Spoken	
[lɪs]	1445	1473	916	416
[lɪʃ]	707	695	826	362
[ləs]	617	612	674	270
[ləʃ]	12	11	14	0

However, if the TP context is extended to include the preceding two segments, we now make the wrong prediction about Pitt and McQueen's Experiments 1-3, since now [dʒu\_] and [bʊ\_] have 100% TP biases towards [s] and [ʃ] respectively. This should have produced a TP effect, but did not. Even worse, the [dɪ\_] and [mi\_] contexts, which produced a large effect in their Experiment 3, are unbiased.

Table 3.14. Triphone frequencies for the stimuli of McQueen and Pitt (1998).

3-phone	Frequency per million words, Celex EFW.CD/EPW.CD			Frequency per million words, Francis/Kucera
	Combined	Written	Spoken	
[dzus]	28	30	4	17
[dzuf]	0	0	0	0
[bus]	0	0	0	0
[buf]	74	79	7	24
[dis]	0	0	0	0
[dif]	0	0	0	0
[neis]	14	15	6	2
[neif]	543	558	394	486
[mis]	0	0	0	0
[mif]	0	0	0	0

A context size of 1 segment is too small to account for the Elman & McClelland (1988) results. A context size of 2 segments can handle those, but not the Pitt & McQueen (1998) results. Larger contexts do not solve this latter problem (since the *juice/bush* stimuli are only three segments long), and in any case lead to a duplication of the lexicon at a prelexical level.

TRACE can explain this disparity, as noted in this connection by Samuel (2000). Lexical effects in TRACE increase over time, and are greatest at the end of long words, because the word nodes take time to reach activation and are more active the more phoneme nodes are feeding into them. The ambiguous fricatives in both experiments came at the end of a word, but the McClelland & Elman words were much longer than the Pitt & McQueen stimuli ([fulɪ\_] and [ɪɹɪkɪjʊlə\_] versus [dʒu\_] and [bu\_]).

In this experiment, Pitt & McQueen not only failed to find a lexical effect with the contexts *jui\_* and *bu\_*, they succeeded in getting a TP effect with the contexts *mee\_* and *nay\_*, both of which make nonwords no matter which way the ambiguous fricative is interpreted. Tables 3.15 and 3.16 show the cohorts at the time the ambiguous fricative appears. It is clear that effects were found in all and only those cases where, in at least one of the paired stimulus contexts, the active cohort strongly favored [s] or [ʃ] at the time the ambiguous fricative appeared.

Table 3.15. Cohorts at the appearance of the ambiguous fricative in the experiment of McClelland and Elman (1988, Experiment 3).

Preceding context		Continued with [s]		Continued with [ʃ]	
		Words	Frequency	Words	Frequency
ridiculou_	[ɪɹɪkɪjʊlə_]	ridiculous	36	(none)	0
fooli_	[fulɪ_]	(none)	0	foolish	11
				foolishly	2

Table 3.16. Cohorts at the appearance of the ambiguous fricative in the experiment of Pitt and McQueen (1998, Experiment 3).

Preceding context		Continued with [s]		Continued with [ʃ]	
		Words	Frequency	Words	Frequency
juı_	[dʒu_]	juice	2	–	0
		juicy	2		
bu_	[bu_]	–	0	bush	4
				bushels	2
				bushes	1
mee_	[mi_]	–	0	–	0
nay_	[nei_]	–	0	nation	45
				nations	43
				nationwide	3
				nationwide	1

No matter what context we choose, there is empirical data which the TP theory will not cover. Our choice of which version to test will have to be based on other grounds. There are two good theoretical reasons to choose a one-segment context for the present study, and a third practical reason. First, we hope to equate "zero-frequency" and "phonotactically illegal". This is plausible for sequences of length 2, but not for those of length 4 – in the latter case, it leads to the claim that any 4-segment word which does not already exist, such as [ʊlɔʃ], is illegal. Second, the MERGE TP theory contrasts with

TRACE in excluding the lexicon from prelexical phonetic processing. As the size of the context increases, so does the number of phoneme sequences whose frequencies the prelexical module has to keep track of. Their number quickly approaches the size of the lexicon. Finally, as a practical matter, long contexts make the frequency counts harder to replicate, since each count is based on a smaller sample.

There is also empirical evidence supporting the one-segment context theory.. Pitt (1988) undertook a replication and extension of the Massaro-Cohen (1983) experiments. He presented an [ɹ]–[l] continuum to American English listeners in the synthetic contexts [d\_æ], [g\_æ], [t\_æ], [b\_æ], and [s\_æ], and measured listeners' "r" and "l" judgments. He found a strong "r" report bias (compared to the baseline [b\_æ]) in [t\_æ], a weaker one in [d\_æ], none in [g\_æ], and a strong "l" bias in [s\_æ].

Absolute per-million frequencies of [l] and [ɹ] after each of the initial consonants are shown in (3.17). The ratio of these yields the *a priori* likelihood that an unknown liquid in that context will be an [l]. As we saw in §3.3.2.3, it is this ratio which, when the listener uses an optimal guessing strategy, predicts the size of the response bias. The order of effects predicted by the likelihood ratio is exactly the order of effects found by Pitt<sup>14</sup>:

---

<sup>14</sup> Pitt himself interpreted these results as *contradicting* the probabilistic account of the phonotactic effect. This is because he assumed listeners were using a suboptimal guessing strategy. Rather than the likelihood ratio, he took the predictor of statistically-induced bias in favor of a given cluster to be the sum of the logarithm of the individual frequencies of the words in which it occurs.

Table 3.17. Likelihood ratio as a predictor of the phonotactic bias effects of Pitt (1998).

Sequence	Frequency (Francis- Kucera)	Ratio, $F([l])/F([ɹ])$	Statistical bias	Empirically measured bias (Pitt 1998)
[tl]	220	0.026	Strong [ɹ]	Strong [ɹ]
[tɹ]	8468			
[dl]	275	0.136	[ɹ]	Weak [ɹ]
[dɹ]	2020			
[gl]	626	0.163	[ɹ]	None
[gɹ]	3845			
[bl]	2407	0.992	None	(Baseline)
[bɹ]	2426			
[sl]	815	62.7	Strong [l]	Strong [l]
[sɹ]	13			

For the present study, therefore, we will pursue a probabilistic theory of phonotactics in perception which makes the following claims:

1. The mechanisms of speech perception have access to a table of length-2 or length-3 sequences occurring in the English lexicon, including their empirical frequencies. We will estimate those from the Celex statistics on British English, checking them against the Francis-Kucera statistics on American English, using the procedure described in the Appendix to this chapter.

2. When an acoustically ambiguous segment between x and y is presented in the context ...A\_B..., it will tend to be parsed as the one which is more frequent in that context. The difference in rate of "x" report between the context A\_B and the context C\_D will depend on the relative likelihood of x and y in those contexts. The influence of statistics will be greatest where the acoustic ambiguity is greatest.

3. The relative likelihood of x and y in A\_B and C\_D can be computed in either of two ways.

a. (INC-1 theory):  $\Pr(x | A_) * \Pr(x | _B)$  compared to  $\Pr(y | A_) * \Pr(y | _B)$ .

b. (SC-1 theory):  $\Pr(x | A_B)$  compared to  $\Pr(y | A_B)$ .

4. The TP effect happens very early, certainly prelexically. However, tasks that directly tap phoneme perception (such as the syllable and phoneme judgment tasks used by Massaro & Cohen (1983)) can be responded to on the basis of either a prelexical phonetic representation, or on the basis of one retrieved from the lexicon after word recognition, following the MERGE proposal of Norris et al. (2000).

### **3.4. A grammar-based account**

A final possibility is that phonotactic regularities are not emergent, but fundamental; that the mechanisms of speech perception have access to the possible, as well as the actual, phonological configurations of their language, and are able to apply that knowledge in perceptual tasks to constrain the hypothesis space.

The chief point at issue between the TRACE and MERGE TP theories on the one hand and a grammatical theory on the other is the status of zero-frequency phoneme sequences. TRACE and MERGE TP treat all such gaps alike: The model simply notes the non-occurrence of a particular sequence, and favors occurring sequences over it. A grammar-based theory can draw the distinction, discussed in §2.1, between true phonological gaps (configurations which cannot occur) and mere lexical gaps (configurations which happen not to have occurred).

That there is such a difference is the central claim we will test here. There are many different ways to implement a theory of grammar in speech perception. The main model parameters are the specific grammar to be used (§3.4.1.) and the rule for using it to decide between alternative interpretations of an ambiguous phoneme (§3.4.2.). The model presented here uses the grammatical framework of phonological Optimality Theory (Prince & Smolensky 1993), entailing a decision model in which multiple candidate parses are entertained in parallel.

### **3.4.1. Choice of grammatical theory**

#### **3.4.1.1. Grammatical framework**

A grammar-based theory of phoneme perception could in principle be built around any procedure which correctly separates the productive gaps from the non-productive ones. I have chosen phonological Optimality Theory (Prince & Smolensky 1993, McCarthy & Prince 1995).

OT is particularly well-suited to phonotactic modelling because phonotactic markedness is a theoretical primitive in OT, embodied in the ranked markedness constraints. Surface representations can be compared for markedness by scoring them with respect to those constraints. This contrasts with rule-based theories, in which markedness is an emergent phenomenon.

Markedness, however, is not the same thing as illegality. A configuration is illegal in an OT grammar if it is never in the output, regardless of what the input is. This is reflected in the grammar by having a markedness constraint against the illegal configuration dominate all of the faithfulness constraints which aim to preserve it, so that an input containing the illegal configuration will be realized without it. The grammar may contain many markedness constraints which do not dominate the relevant faithfulness constraints and hence do not trigger repairs; configurations violating only such constraints are not illegal (though they may be marked in other ways).

An example is the case of \*PAL, the constraint which forbids palatal consonants (§2.3.2.3.4). It dominates the faithfulness constraint IDENT[BACK], and hence is able to compel violations of it:

(3.18) \*PAL » IDENT[BACK]

/ca/	*PAL	IDENT[BACK]
[ca]	*!	
□ [ka]		*

A constraint C in a grammar is said to be *active* for an input *i* if at least one candidate is eliminated by C (Prince & Smolensky 1993, Chapter 5). In (3.18), for example, \*PAL is active for the input /ca/, because it is there that [ca] is eliminated.

In a given grammar, some constraints are never active for any input. An example in English is \*VOICE]□, forbidding voiced obstruents in syllable codas (McCarthy 1998). Voicelessness is obligatory for coda obstruents in many languages, including the standard varieties of Russian, Polish, German, and Turkish. Illegal coda clusters are repaired phonologically by devoicing, indicating that faithfulness constraints are being violated in order to satisfy \*VOICE]□. (Hence, \*VOICE]□ is active for some inputs in those languages – it is the constraint which eliminates the candidate outputs with voiced coda obstruents.)

In English, however, \*VOICE]□ is ranked too low to have such an effect. English tolerates voiced coda obstruents (*tub, leave, brag, etc.*). No candidate is ever eliminated by \*VOICE]□, which, therefore, is *inactive* in English.

(3.19) IDENT (for instance) » \*VOICE]□

/liv/	IDENT	*VOICE]□
□ [liv]		*
[lif]	*!	

The general unmarkedness of voiceless codas in acquisition and cross-linguistically, like the general markedness of palatals, are crucial to any grammatical model of language, but are outside the scope of statistically-based models such as TRACE and MERGE TP.

### 3.4.1.2. Particular grammar

Within a given framework, there are at least as many different grammars as there are languages, and for each language, perhaps as many as there are linguists. The predictions of the perceptual model depend on which one is selected. If experiment falsifies these predictions, the problem may lie with either the perceptual theory itself or with the grammatical analysis of the given language (just as, if experiment falsifies a probabilistic theory, the problem may be due to the perceptual theory itself, or to faulty frequency counts).

It is therefore important to start by examining phenomena whose predicted perceptual effects are insensitive to the choice of a specific analysis. The lack of onset [tld] clusters, for instance, is a good choice, because those onsets are so robustly illegal that any theory of English grammar (Optimality-Theoretic or not) has to ban them. In an OT framework, that means they must be ruled out by *some* markedness constraint, which *ipso facto* is active. Even if our analysis in Chapter 2 has pointed the finger at the wrong markedness constraint, any alternative analysis will have a different one from which the same perceptual consequences will follow.

Thus, we expect the perceptually influential phonotactic constraints of a language to include, at a minimum, the ones which are by all measures productive: those that are not naturally violated (having no lexical exceptions) and that speakers cannot be induced to violate (either because the banned configurations trigger repairs, or because the speaker simply cannot pronounce them without great effort). For more detailed discussion, see §2.2.

### 3.4.2. Decision mechanism

The most straightforward adaptation of OT to speech perception is essentially this: Linguistic effects on speech perception come about because language limits the set of available parses. The listener constructs a phonological parse at two levels of representation, corresponding to OT's underlying /UR/ and surface [SR]. The [SR] is computed from the acoustic signal, while the /UR/ is retrieved from the lexicon. Different (/UR/, [SR]) pairs compete to account for the observed signal, with the grammar as referee: The (/UR/, [SR]) pairs are scored by the hierarchy of markedness and faithfulness constraints of the language, and perception favors the most harmonic pair. Thus, the OT grammar does the same job in speech perception that it does in linguistic theory: It compares (/UR/, [SR]) pairs and picks the most harmonic.

For example, in Pitt (1998)'s replication of the experiments of Massaro & Cohen (1983), using nonword stimuli, there are no /UR/s to deal with, so the issue is decided by the markedness constraints:

(3.20) Both endpoints legal □ no grammatical bias

UR = •	OCP(CONT, COR)	SPREAD [COR]
a. (•, [bɪæ])		
b. (•, [blæ])		

(3.21) [l] illegal □ [ɹ] bias

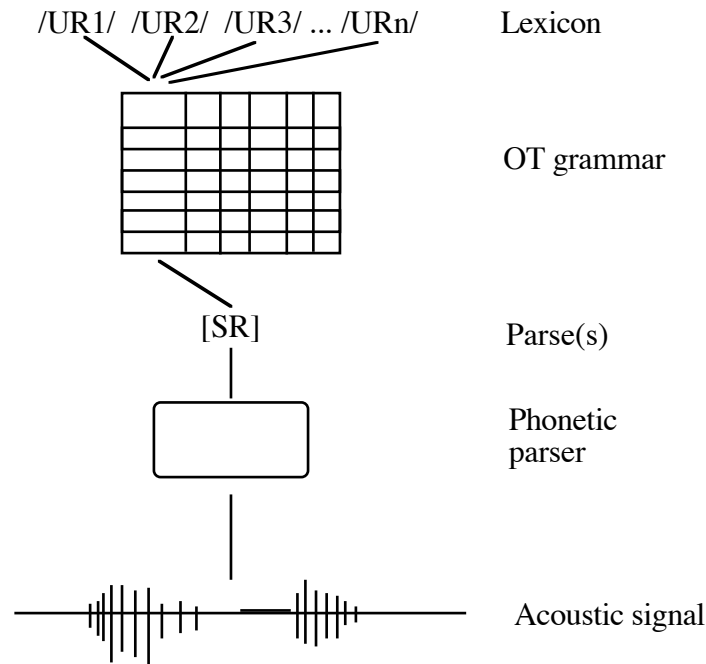
UR = •	OCP(CONT, COR)	SPREAD [COR]
a. (•, [tɹæ])		
b. (•, [tlæ])	*!	

(3.22) [ɹ] illegal □ [l] bias

UR = •	OCP(CONT, COR)	SPREAD [COR]
a. (•, [sɹæ])		*!
b. (•, [slæ])		

In this model, the incoming acoustic signal is first transduced into one or more surface phonetic representations, or [SR]s. The transducing mechanism is a black-box component which in Figure 3.3 is labelled "Phonetic Parser"; it could also be called a "Feature Extractor". Given a speech stimulus, it produces a set of [SR]s consistent with that stimulus.

Figure 3.23. Architecture of an OT-grammatical-based parsing model.



I assume that under normal laboratory conditions, with a short stimulus clearly spoken, the Phonetic Parser will emit a single [SR]. Two or more [SR]s can be coaxed out of it by presenting an acoustically ambiguous stimulus. The likelihood that a stimulus between, say, [r] and [l] will evoke [r] is assumed to be independent of the likelihood that it will evoke [l]; for a given stimulus level, there is a certain probability of getting [r], a certain probability of getting [l], a certain probability of getting both, and a certain probability of getting neither. These probabilities change depending on the acoustic constitution of the stimulus.

The candidate [SR]s are assumed to represent syllabification. It is not in dispute that syllables can be incorporated into a prelexical representation, since nonsense words, which lack a lexical representation, can be syllabified in off-line judgment tasks. The question is whether the syllabic structure is automatically computed as part of the parsing process. There is evidence that it is. Syllable boundaries are needed for segmentation and lexical access, so they have to be marked in the input to the lexical-access stage. In on-line

word-spotting tasks, English listeners are better at finding a word boundary when it is aligned with the left-hand boundary of a stressed syllable (McQueen et al. 1994, Cutler & Norris 1988), which suggests that the input is parsed exhaustively in tofeet. A word boundary is harder to find if a syllable boundary drawn at that point would create a phonotactically impossible syllable (e.g., spotting *apple* in *fapple*) (Norris et al. 1997, McQueen et al. 1998). Abstract grammatical preferences for maximal onsets, and for syllabifying intervocalic consonants with the more stressed vowel, are also detectable in word-spotting – generalizations which only make sense when stated in terms of syllables (Kirk 2001).<sup>15</sup>

As in the Shortlist model of Norris (1994), lexical entries are activated by an [SR] which is sufficiently similar to them. I will assume that "sufficient" similarity is determined by the neighborhood metric; a lexical competitor is any lexical item whose /UR/ can be obtained from one of the active [SR]s by a one-segment insertion, deletion, or replacement. Competition between all of the active (/UR/, [SR]) pairs then takes place through the grammar.<sup>16</sup>

It is at this point natural to posit that the grammar scores each pair and chooses the most harmonic. This theory is elegant but fatally flawed. The root of the problem is that such a literal implementation of OT inherits the OT principle called "Strictness of

---

<sup>15</sup> Evidence against syllabification comes chiefly from sequence-monitoring tasks. These are known to be sensitive to syllabification in certain languages. For example, Mehler, Dommergues, Frauenfelder, and Segui (1981) found that French listeners detected a CV or CVC target faster when it exactly matched the first syllable of the stimulus word – *ba* being detected faster in *ba.lance* than in *bal.con*, but *bal* being detected faster in *bal.con* than in *ba.lance*. English speakers show no such difference, whether tested with English (*ba/bal* in *bal.cony* and the ambisyllabic *ba.lance/bal.ance*) or with the original French materials (Cutler, Mehler, Norris, & Segui 1986). The authors interpreted this to mean that English listeners do not use on-line syllabification to segment speech, even in cases like *bal.cony* where the syllabification is unambiguous. However, Kirk (2001) argues instead that, owing to the effects of stress on syllabification, the first syllable of *balcony* is *balc*, and hence that neither target matched a syllable. For an extensive critical review of the evidence for and against on-line syllabification in English and other languages, see Kirk (2001, Chapter 2).

<sup>16</sup> Although this description is confined to the single-word stimuli actually used in the experiments, it can be extended to longer utterances in a straightforward way. The Phonetic Parser emits one or more candidate [SR]s, as before. Candidate word /UR/s are activated by sufficiently similar substrings of the [SR]s; candidate utterance /UR/s are the concatenations of nonoverlapping word /UR/s. These utterance (/UR/, [SR]) pairs then compete as before. This allows the theory to capture word segmentation and inter-word sandhi phenomena.

Domination", which states that if Constraint A is ranked above Constraint B, then violating Constraint A even once is less harmonic than violating Constraint B any number of times. Strict domination is the only means which OT affords to represent one constraint's primacy over another. Violating A cannot be just a little bit worse than violating B; it must be either infinitely worse, equally bad, or infinitely better (Prince & Smolensky 1993:78).

One consequent prediction is that inactive markedness constraints should have just as big an impact on perception as active ones. In order to account for cross-linguistic patterns in the sound inventories of languages, OT posits certain universally fixed rankings – e.g., that labial and dorsal articulations are universally more marked than coronal articulations. In the grammar of English, which allows both labials and coronals, neither the anti-labial constraint nor the anti-coronal constraint is active; however, they are still in the grammar and one still dominates the other. A stimulus which is ambiguous between [ba] and [da] therefore has two interpretations, one of which violates the constraint against labials, the other of which violates the lower-ranked constraint against coronals. Since labials are by hypothesis infinitely worse than coronals, perception should strongly favor [da] over [ba]. This does not seem to always be the case: For example, when Luce (1986, Ch. 3) presented listeners with a balanced set of CVC nonsense words in noise at a +5 dB signal-to-noise ratio, he found that final [b] was reported as [d] 27 times out of 150, while final [d] was reported as [b] 24 times.<sup>17</sup>

In order to make the necessary distinction between significant and insignificant markedness differences, we must stipulate that *inactive* markedness constraints carry little if any perceptual weight. That is, the (/UR/, [SR]) pairs are evaluated only by that part of the constraint hierarchy which ranks above the highest-ranked inactive markedness constraint.

---

<sup>17</sup> Syllable-initially, [b] was reported as [d] far more often than the reverse: 39 times versus 4 at a +5 dB signal-to-noise ratio, 23 times versus 4 at –5 dB. Even syllable-final [b] was reported as [d] 15 times at a –5 dB ratio, versus 6 times the other way around. It may be that the low markedness, or perhaps the high frequency, of [d] is having some sort of effect. This effect is, however, not as overwhelming as expected, and may in any case be due to the spectral quality of the noise used (white Gaussian noise up to 4.8 kHz), which is similar to the diffuse-rising spectrum of alveolar plosive bursts (Blumstein & Stevens 1979).

A second problem that this proposal runs into is the same one which bedeviled TRACE: the inability to adjust the influence of the lexicon based on attentional factors. Lexical effects, such as the Ganong effect, would in this theory be captured by faithfulness constraints:

(3.24) [d] nonword, [t] word □ [t] bias

UR = /tæsk/	ID-VOICE
a. (/tæsk/, [dæsk])	*!
b. □ (/tæsk/, [tæsk])	

(3.25) [d] word, [t] nonword □ [d] bias

UR = /dæʃ/	ID-VOICE
a. □ (/dæʃ/, [dæʃ])	
b. (/dæʃ/, [tæʃ])	*!

The violation, and hence the predicted bias, is just as large regardless of how much attention the listener allocates to the lexicon, contrary to the findings of Cutler et al. (1987).

This is only an apparent problem. All experimental results are obtained by averaging over a large number of trials. Suppose that on each trial, the listener either "attends to the lexicon" – i.e., insists on a parse with an /SR/ – or does not, on a trial-by-trial basis. The task manipulations used by Cutler et al. (1987) can be seen as changing the probability, rather than the extent, of the listener's attention to the lexicon on each trial, and hence the number of lexically-biased responses which were averaged into the data.

The decision rule we must adopt is therefore:

(3.26)

- a. If attending to the lexicon, and if the stimulus is close enough to a real word to activate some /SR/s, choose the (/UR/, [SR]) pair which scores best on the active constraints.
- b. Otherwise, choose the ([SR]) which scores best on the active markedness constraints (the faithfulness constraints have no function in the absence of a /UR/).
- c. Ties are broken randomly.

As in the Race model of Cutler et al. (1987), responses to phoneme tasks can be based on either the computed [SR] or the retrieved /UR/, with task constraints dictating which is favored in each case. There is laboratory evidence for the existence of both levels and for attentional effects on them.

Xu (1991) showed that Mandarin Chinese speakers had poorer recall for written lists of rhyming morphemes when the list elements shared the same tone than when they differed in tone. Speakers were then asked to perform the same task with lists constructed so that the first two items had the same surface tone but different underlying tones, and performance was compared with lists in which the first two items had different surface and underlying tones. Performance was worse on lists of the first sort, suggesting that the short-term memory representation in this task was in terms of [SR]s.

On the other hand, Lahiri & Marslen-Wilson (1991), using a gating task, found that listeners interpreted vowel nasalization differently depending on their native language: English listeners took it as a sign of an upcoming nasal consonant, since English vowels are not inherently nasalized, but become so in a nasal phonetic context. Bengali listeners, on the other hand, speak a language which has both inherently (i.e., contrastively) nasalized vowels and contextually nasalized vowels. They overwhelmingly interpreted vowel nasalization as underlying (i.e., did not take it as a sign that a nasal consonant was coming up) until they actually heard the beginnings of the nasal consonant. This suggests that the Bengali lexicon represents contextually nasalized vowels as not nasalized, showing a difference between lexical representation and surface phonetic representation which the gating task (an inherently lexical task) revealed. It further indicates that the Bengali speakers were

choosing the more faithful (/UR/, [SR]) pair in which the underlying and surface vowel had the same degree of nasalization over the less faithful pair in which they differed, as we would expect.<sup>18</sup>

### 3.5. Summary

This chapter has presented three very different theories of phonotactics in speech perception.

The TRACE model sees phonotactics as an effect of similarity to lexical items. Of the three theories, it makes the smallest demands on the learner, requiring knowledge only of the lexicon. Phonotactic effects are viewed as diluted lexical effects, in which permitted configurations are supported by partially-overlapping lexical items, which allows them to defeat competing illegal candidates via lateral inhibition at the phoneme level.

The MERGE TP model attributes phonotactic effects to differing frequencies of short phoneme sequences. The theory requires knowledge of the lexicon and of a set of attested phoneme sequences, which may be quite large but can be acquired straightforwardly through observation. Phonotactic effects are taken to occur at a pre-lexical level, with rare sequences being perceptually disfavored.

The OT grammatical model sees perceptual phonotactic effects as a consequence of the limited range of parses available in the language, and the listener's bias towards a parsed percept. The implementation used here requires knowledge of the lexicon, and of a set of

---

<sup>18</sup> If the OT interpretation of Lahiri & Marslen-Wilson's results is correct, it is empirical evidence against the OT principle of Lexicon Optimization (Prince & Smolensky 1993, Inkelas 1994). Lexicon Optimization is a means of dealing with the source-filter nature of the OT grammatical model, which can map several /UR/s to the same [SR]. In acquiring the lexicon, it is asserted, the /UR/ which is chosen is the one to which the observed [SR] is most faithful.

We would be led to expect Bengali speakers to represent surface [CV<sup>n</sup>N] words as underlying /CV<sup>n</sup>N/, which map to the same output more faithfully than an underlying /CVN/ would. We would therefore expect a gating stimulus of the form [CV<sup>n</sup>...] to often be completed with an N, i.e., matched to a word whose underlying representation is /CV<sup>n</sup>N/. Instead, they were overwhelmingly matched to words whose underlying representation was /CV<sup>n</sup>C/, suggesting that the surface [CV<sup>n</sup>N] words are underlyingly /CVN/. The study's finding that speakers apparently lexicalize surface contextually nasalized vowels as underlying non-nasalized vowels indicates they are not using Lexicon Optimization.

constraints. The number of constraints needed is probably not very large (a grammar of the syllable onsets of English, in Chapter 2, needed well under 20), and the correct ranking is provably learnable (Tesar & Smolensky 1995); however, their provenance is unclear. They are normally taken to be innate, since the patterns they represent occur world-wide.<sup>19</sup> Phonotactic effects are assumed to occur at a prelexical level, the level of surface representations, with banned sequences being perceptually disfavored.

Each of these theories suffers from empirical drawbacks in one domain or another. TRACE has difficulty explaining why phonotactic effects are more robust than lexical effects. The MERGE TP model cannot be pinned down on precisely which phoneme sequences are perceptually relevant; different choices leave different lab results unexplained. The OT grammatical model accounts for effects of illegality, but not the apparent (usually very small, but definitely detectable) effects of sequence frequency or as lexical neighborhood (Newman, Sawusch, & Luce 1997; Pitt & McQueen 1998 Exp. 4).

The drawbacks of one model are, naturally, the advantages of the others. TRACE is theoretically attractive because it offers an extremely parsimonious learning model. Because sound-meaning relations are arbitrary, the lexicon must be learned in any theory. TRACE says that *only* the lexicon must be learned, and that apparent effects of grammatical regularity are really emergent properties of lexical interaction. MERGE TP is only slightly less parsimonious – only the lexicon must be learned, but the relevant regularities have to be actively abstracted from it by the probability-tracking system. Though both theories require innate *structure* in the perceptual system, neither requires detailed innate *knowledge* the way the OT grammatical theory does. As a practical matter, it is also easier to make predictions from TRACE and MERGE TP than from any grammatical theory, since less analytic depth is required.

---

<sup>19</sup> Moreover, since the constraints are violable and do get violated, they cannot be individually inferred from the speech corpus by any simple mechanism—especially the markedness constraints, being prohibitions for which no positive evidence can exist. (Naturally, a linguistically more sophisticated mechanism could take advantage of alternations to deduce abstract underlying forms and the markedness constraints necessary to cause the alternations.)

In CHAPTER 4, our focus will be on the interesting claim, put forth by the TRACE and MERGE TP theories and denied by the OT grammatical theory, that phonotactic illegality is equivalent to zero frequency. The claim is interesting because it suggests that phonology, at least in perception, is considerably simpler than many linguists have hitherto supposed, and offers a means of circumventing the difficult problem of grammar acquisition.

### **3.6. Appendix: Computing frequencies**

All frequency counts were made from the Celex lexical database (Baayen et al. 1995). This is based on a corpus of 16.6 million words of written English and about 800,000 words of spoken English. Most of the corpus is from British sources, and the phonetic transcriptions are British.

Celex provides two ASCII phonetic transcription systems. I used the one found in Field 7 of the file EPW.CD. Variant pronunciations are given for some words, but I always used only the first pronunciation listed.

Celex gives frequency counts by "lemma" (i.e., citation form, with *know*, *knows*, *knew*, and *knowing* all lumped together) and by "wordform" (i.e., counting inflected forms separately). In both cases, homophonous words belonging to different grammatical categories are counted separately (e.g., *link* noun and *link* verb). I used the wordform database (except where otherwise noted), specifically, the files EPW.CD (the pronunciations) and EFW.CD (the frequencies).

Frequencies are counted separately for the written and spoken corpora. A "combined" frequency count is also given; since most of the corpus is written, the "combined" frequency is usually very close to the written frequency. I have used the spoken counts except where otherwise noted (non-spoken frequencies are used only for compatibility with counts based on the Francis-Kucera (1967) written-corpus norms). All

counts are from the Celex per-million-words estimates (combined, Field 6; written, Field 9; spoken, Field 12; of EFW.CD).

The Celex transcription system marks syllable boundaries and includes stress marks for primary and secondary stress. These were removed.

The scripts used to create the counts are appended.

### (3.27) *Script for counting frequency of length-n sequences*

```
#!/usr/local/bin/perl

# make_ngram_table

# usage:  make_ngram_table <n>

# where <n> = # of segments per gram

# Each word is enclosed in wd boundary markers "(" and ")", which
# count as phonemes.

$n = $ARGV[0];

$phon_db = '/tmp/Celex/EPW.CD';
$freq_db = '/tmp/Celex/EFW.CD';
open (PHON, "< $phon_db") || die "Couldn't open $phon_db";
open (FREQ, "< $freq_db") || die "Couldn't open $freq_db";

while ( ($phon_buf = <PHON>) && ($freq_buf = <FREQ>) ) {

    # read a record from EPW.CD and EFW.CD

    ($phon_IDnum, $phon_orth, $phon_freq_comb, $foo, $foo, $foo, $pron) =
        split /\\\/, $phon_buf;

    ($freq_IDnum, $freq_orth, $foo,
     $freq_comb, $freq_comb_dev, $freq_comb_perM, $freq_comb_log10,
     $freq_writ,           $freq_writ_perM, $freq_writ_log10,
     $freq_spok,          $freq_spok_perM, $freq_spok_log10
    ) = split /\\\/, $freq_buf;

    ( ($phon_IDnum == $freq_IDnum) && ($phon_orth eq $freq_orth) ) ||
        die "$phon_db mismatches $freq_db:\n$phon_buf$freq_buf";

    # Purge pronunciation of non-segmental characters
    ($segment_pron = $pron) =~ tr/\-\'\"//d;

    # Find the n-grams and count their frequencies
    @segments = ( '(', (split ' ', $segment_pron), ')' );
    foreach $i (0..($#segments - $n + 1)) {
        $ngram = join ' ', @segments [$i..($i + $n - 1)];
```

```

    $freq_comb_perMs {$ngram} += $freq_comb_perM;
    $freq_writ_perMs {$ngram} += $freq_writ_perM;
    $freq_spok_perMs {$ngram} += $freq_spok_perM;
}
}

# Print out the ngrams and their frequencies
foreach $ngram (keys %freq_comb_perMs) {
    printf "%s\t%5d\t%5d\t%5d\n",
        $ngram,
        $freq_comb_perMs {$ngram},
        $freq_writ_perMs {$ngram},
        $freq_spok_perMs {$ngram};
}

```

(3.28) *Script for turning those counts into TPs*

```

#!/usr/local/bin/perl

# make_TP_table

# Given a list of n-grams and their frequencies, computes transitional
# probabilities from X1X2...X(n-1) to Xn.

# Input format is

# <X1...X(n-1)Xn> <combined freq> <written freq> <spoken freq> etc.

# Output format is

# <X1...X(n-1)> <Xn> <P(Xn | X1...Xn-1), combined> <same, written> etc.

while ($buf = <STDIN>) {

    ($ngram, @freqs) = split /\s+/, $buf;

    @segments = split '', $ngram;
    $lastseg = pop @segments;
    $context = join '', @segments;

    # Count occurrences of each context X1...X(n-1), and of each
    # ngram X1...Xn.
    foreach $i (0..$#freqs) {
        $context_freqs [$i] {"$context"} += $freqs [$i];
        $ngram_freqs [$i] {"$context$lastseg"} += $freqs [$i];
    }
}

# Compute transition probabilities conditional on X1...X(n-1).

foreach $ngram (keys %{$ngram_freqs[0]}) {

    @segments = split '', $ngram;
    $lastseg = pop @segments;
    $context = join '', @segments;

```

```

foreach $i (0..$#freqs) {
    $TPs [$i] = '(none)';
    next unless $context_freqs [$i] {"$context"}; # avoid /0 errors
    $TPs [$i] = sprintf "%6.3f",
        ($ngram_freqs [$i] {"$context$lastseg"} / $context_freqs [$i]
        {"$context"});
    }

    print "$context\t$lastseg\t";
    print join "\t", @TPs;
    print "\n";
}

```

(3.29) *Script for finding the active cohort following a given phonological string*

```

#!/usr/local/bin/perl

# cohort

# Given a phoneme string, find all words in Celex EPW.CD/EFW.CD
# which begin with that string. Print each word and its per-
# million frequencies.

# Usage: cohort <string>

$beginning = shift @ARGV;

open (PHON, "cat /tmp/Celex/EPW.CD |") || die "Couldn't open EPW.CD";
open (FREQ, "cat /tmp/Celex/EFW.CD |") || die "Couldn't open EFW.CD";

while (($phon_buf = <PHON>) && ($freq_buf = <FREQ>)) {

    ($phon_IDnum, $phon_orth, $phon_freq_comb, $foo, $foo, $foo, $pron) =
        split /\//, $phon_buf;

    ($freq_IDnum, $freq_orth, $foo,
     $freq_comb, $freq_comb_dev, $freq_comb_perM, $freq_comb_log10,
     $freq_writ, $freq_writ_perM, $freq_writ_log10,
     $freq_spok, $freq_spok_perM, $freq_spok_log10
    ) = split /\//, $freq_buf;

    ( ($phon_IDnum == $freq_IDnum) && ($phon_orth eq $freq_orth) ) ||
        die "$phon_db mismatches $freq_db:\n$phon_buf$freq_buf";

    # Purge pronunciation of non-segmental characters
    ($segment_pron = $pron) =~ tr/\-\'\"//d;

    # Is it in the cohort?
    next unless ($segment_pron =~ /\^Q$beginning\E/);

    # Yes -- print
    printf "%s ", $beginning;
    printf "%6d ", $freq_comb_perM;
    printf "%6d ", $freq_writ_perM;
    printf "%6d\t", $freq_spok_perM;
    printf "%s\t", $freq_orth;
    printf "%s\n", $segment_pron;
}

```

(3.30) *Program for simulating statistically-based guessing*

```
#!/usr/local/bin/perl

# simulated_guess

# Simulated experiment, illustrating the usefulness of TPs. A subject
# hears a corpus of English (a list of words, each word selected from
# Celex such that the English vocabulary occurs with its empirical
# frequency). At random, infrequent intervals, a word is truncated
# after at least (n-1) segments. The listener predicts the next one
# by consulting a table of n-grams, and guessing that the next segment
# will be whatever is most likely to follow the last (n-1) segments
# of the stimulus.

# Input is output of make_ngram_table, awked to: <gram> <freq>

# Output is the expected proportion of trials on which the subject
# guesses correctly.

# Count frequencies

while ($buf = <STDIN>) {
    ($gram, $freq) = split /\s+/, $buf;

    @segs = split '', $gram;
    $nextone = pop @segs;
    $context = join '', @segs;

    $totfreq      += $freq;      # frequency with which ngrams occur
    $cfreq {$context} += $freq;  # freq of ngrams starting with this (n-1)
gram
    $gfreq {$gram}    += $freq;  # frequency of this ngram

    # Keep track of likeliest ngram beginning with each (n-1)gram
    $current_best = $best_guess {$context};
    if ($freq > $gfreq {"$context$current_best"}) {
        $best_guess {$context} = $nextone;
    }
}

$n = length ($gram);

# Print guessing strategy
foreach $context (keys %cfreq) {
    print "$context $best_guess{$context}\n";
}
print "\n";

# Simulate experiment

srand (time());
$CELEX_SIZE = 1800000;
$TRIALS = 100000;

open (EPW, "cat /tmp/Celex/EPW.CD | ") || die "Couldn't open EPW";
```

```

while ($buf = <EPW>) {
    ($foo, $orth, $comb_freq, $foo, $foo, $foo, $pron) = split /\//, $buf;

    next unless $comb_freq;

    $seg_pron = '';

    TRIAL: for ($i = 1; $i <= $comb_freq; $i++) {

        # Each word gets as many lottery tickets as its frequency,
        # and each ticket has $TRIALS/$CELEX_SIZE chance to win. This
        # insures that any given word has its natural probability of
        # being used on any given trial, and that the expected # of
        # trials is $TRIALS.

        $r = int (rand ($CELEX_SIZE));
        if ( $r >= ($CELEX_SIZE - $TRIALS)) {

            unless ($seg_pron) {
                ($seg_pron = $pron) =~ tr/\'\\"-\//d;
                $seg_pron = "(" . $seg_pron . ")";
                last TRIAL if (length ($seg_pron) < $n);
            }

            print "$orth $i $seg_pron ";

            $ngram_start = int (rand (length ($seg_pron) - $n + 1));
            $ngram = substr ($seg_pron, $ngram_start, $n);
            $context = substr ($ngram, 0, $n-1);
            $nextone = substr ($ngram, -1);
            print "$context $nextone\n";
            $total_trials++;
            $correct_trials++ if ($nextone eq $best_guess {$context});
        }

    }

}

print "$n: $correct_trials right out of $total_trials: ";
printf "%6.3f\n", $correct_trials/$total_trials;

```