Chapter 4

Top-down lexical factors

4.1 Introduction

This chapter aims to examine the top-down lexical factors that play a role in naturalistic misperception. Some of the previous analyses of naturalistic misperception using the sub-corpora of the our combined mega corpus have identified a few top-down lexical factors such as segmental frequency (Bird, 1998), syllable factors (Browman, 1978) and word frequency (Bond, 1999; Vitevitch, 2002; Tang and Nevins, 2014). Overall, their findings were encouraging, suggesting that top-down lexical factors do have an effect on naturalistic misperception, and that they are consistent with experimental findings. However, given that the data were collected by different people, it is possible that their findings are susceptible to certain idiosyncrasies due to reporting biases. Furthermore, the amount of naturalistic data used by these studies was small; therefore, it is possible that their findings are due to chance. These drawbacks highlight the need for a reanalysis of these findings using the combined corpus. In addition to the three top-down factors, mentioned above (segmental frequency, syllable factors, and word frequency), the effect of the conditional probability of a word in an utterance was also examined (which, in information theory, is referred

to as self-information, Shannon, 1948). This chapter will examine whether there are top-down effects from linguistic units of various sizes – segments, syllables, words, and utterances. If so, how strong are these effects? These four factors serve as four main sections in this chapter. Each of the four sections is introduced below, starting with segmental frequency.

4.1.1 Segmental frequency

The role of segmental frequencies in segmental confusions will first be examined. Segmental frequency is the frequency of the occurrence of segments found in a large sample of the language. I selected two aspects of segmental confusions that could be explained with segmental frequency.

4.1.1.1 Target and response biases

The first aspect of segmental confusions concerns the target and response biases in misperception. Target bias means that certain phones are more (or less) likely to be spoken but misperceived. That is, given there is a misperception, not all segments are equally likely to be the target. Similarly, response bias means that certain phones are more (or less) likely to be the resultant perceived phones in a misperception. That is, given there is a misperception, not all segments are equally likely to be the response. Can the biases (if any) be explained by the segmental frequency in the language?

It is important to understand that the target bias means that certain phones are more likely to surface as the target of a misperception, and it does *not* mean that certain phones are more likely to undergo misperception. In other words, target bias is referring to the probability of a phone being the target segment of a misperception, and it is *not* referring to the probability of a phone being erroneously misperceived. To further clarify what target and response biases are, let us consider the following

example. 100 phones were presented to a listener and 40 phones were misperceived. Amongst these 40 phones (the intended segments), what is the distribution of the intended segments? Can the distribution of the intended segments be predicted by the distribution of their segmental frequency in the language? These 40 phones were misperceived as another 40 phones (the perceived segments). What is the distribution of the perceived segments? Can the distribution of the perceived segments be predicted by the distribution of their segmental frequency in the language?

To examine this, the frequency of being a target in a misperception will be computed for each segment type. This will then be correlated with the frequency of each segment type found in the language. A significant correlation would mean that when a segment is misperceived, the likelihood of this segment being segment x is dependent on how frequent segment x is in the language. Similarly, for the response bias, the frequency of being a response in a misperception will be computed for each segment type, and will then be correlated with the frequency of each segment type found in the language. A significant correlation would mean that when a segment is misperceived, the likelihood of the perceived segment being segment x is dependent on how frequent segment x is in the language.

4.1.1.2 Asymmetrical confusion

The second aspect of segmental confusions concerns their asymmetrical patterns. In Chapter 3, Section 3.8, three well-known asymmetrical patterns in English were analysed, namely TH-fronting, velar nasal fronting, and back vowel fronting. The question is whether asymmetrical patterns such as these ones, and in general, can be predicted by the relative frequencies of the segments in the language. For instance, [θ] is being perceived as [f] more often than the reverse, but it is also true than [f] is a more frequent segment than [θ]. If the relative frequencies can affect the asymmetrical confusions, then confusion asymmetries are a function of both perceptual biases and

frequency biases.

4.1.1.3 Frequency measures

Furthermore, three different measures of segmental frequency will be examined for the strength of their effect on the three aspects of segmental confusions mentioned above. The three measures are token frequency (the number of times a given segment is found in the language), type frequency (the number of words that contain a given segment) and weighted type frequency (the number of words that contain a given segment, weighted by the token frequency of the words).

The unweighted type frequency measure is purely lexically-based, while the token frequency measure is not. The weighted type frequency measure is a hybrid measure, which is partially lexically-based. If we find that the type frequency measure generally predicts segmental confusions better than the other two non-lexically-based measures, then we could argue that listeners are sensitive to lexical items in segmental misperception.

4.1.2 Syllable factors

Moving away from segments into syllables, we could examine whether certain factors on the syllable level have a top-down effect on segmental misperception. Three factors are tested – syllable constituency, syllable position and stress.

Syllable constituency is the position of the segment in a syllable, namely onset, nucleus and coda. Syllable position is the position of the syllable that contains the segment in a polysyllabic word. Three positions can be generalised, namely word initial, word medial, and word final. Stress is whether the syllable is stressed or unstressed.

Focusing on whether these factors have an effect on whether a segment is more likely to be misperceived, the following questions could be asked: Do segmental

errors occur evenly across the three syllable constituents? Do we expect segments in certain syllable positions to be misperceived more often than others? Are segments in unstressed syllables more often misperceived than those in stressed syllables? Finally, do we expect the effect of syllable constituency and stress to be different between monosyllabic and polysyllabic words?

4.1.3 Word frequency

Let us move on to a larger linguistic unit. The relationship between the frequency of the intended word and that of the perceived word will be examined. First, is there a relationship between the frequency of the intended word and that of the perceived word, $Freq._{Perceived} = f(Freq._{Intended})$? Second, is the frequency of the perceived word more frequent than or similar to that of the intended word, $Freq._{Perceived} > or \approx Freq._{Intended}$?

This will allow us to find out whether listeners are sensitive to frequency (or its correlates) on the segmental level as well as the word level in misperception. Furthermore, it will shed light on the mechanisms/strategies that listeners use when retrieving lexical items in speech perception.

4.1.4 Self-information

The last top-down factor concerns the amount of self-information a word has and its effect on whether a word is more likely to be misperceived. By self-information, we are referring to Shannon information, which is a function of the average unpredictability in a random variable (Shannon, 1948).

Two kinds of self-information were tested. One is based on the unconditional probability of a word, which is basically the token frequency of a word in a sample of the language. The other is based on the conditional probability of a word, given its previous words. The self-information of a word is the negative log of the probability

of a word; therefore, the more probable a word is, the less self-information it has.

Our question is whether the amount of self-information of a word can be used to predict how likely it is that it will be misperceived in an utterance. If the conditional self-information is shown to be a good predictor after taking into account the unconditional self-information, then it would show that listeners are sensitive, not only to the token frequency of a word, not also to the frequency of a word given its context.

Furthermore, the direction of the effect of self-information on the likelihood of word errors can inform us of possible causes of misperception. On the one hand, it is well-known that high frequency words have a lower processing cost than low frequency words (Brysbaert and New, 2009; New et al., 2007; Keuleers, Brysbaert, and New, 2010; Ernestus and Cutler, 2014). Therefore one possible cause of misperception is that words with high self-information (therefore low frequency/less probable) are more likely to be misperceived because of processing difficulties. On the other hand, words with low self-information (therefore high frequency/more probable) are prone to phonetic reduction (Wright, 1979; Aylett and Turk, 2004; Bybee, 1995; Bybee and Hopper, 2001; Bybee, 2001; Coetzee and Kawahara, 2013). Therefore, words with low self-information are more likely to be misperceived because of the amount of phonetic information. Since the two explanations make different predictions, our analyses can reveal which of the two is more plausible.

4.1.5 Summary

This chapter is broken down into five sections. The first four sections contain the analyses of the four top-down factors described previously. First, Section 4.2 will examine the effect of segmental frequency on two different aspects of segmental confusions. Second, Section 4.3 will evaluate the effect of three syllable factors on the likelihood of a segment error. Third, Section 4.4 will examine the frequency

relationship between the intended and perceived words. Fourth, Section 4.5 will evaluate the effect of self-information on the likelihood of a word error in an utterance. Finally, Section 4.6 will conclude the findings and contributions made in this chapter.

4.2 Segmental frequency

This section examines the role of segmental frequencies in segmental confusions. Segmental frequency is the frequency of the occurrence of segments observed in a representative sample of the language. We will focus on two aspects of segmental confusions that could be the result of the segmental frequencies in the language. In addition, three different frequency measures will be tested.

The first and the simplest measure is token frequency, which is the number of occurrences of a given phone (Kučera and Francis, 1967; Nusbaum, Pisoni, and Davis, 1984). The second measure is type frequency, which is the number of lexical items containing a given phone (Kučera and Francis, 1967; Nusbaum, Pisoni, and Davis, 1984). The third measure is like the second measure but weighted by the token frequency of each of the lexical items containing a given phone (Nusbaum, Pisoni, and Davis, 1984). They are summarised below.

- Token: The number of occurrences of a given phone
- Type: The number of lexical items containing a given phone
- Type (Weighted): The sum of the log-transformed frequencies of the lexical items containing a given phone

The role of frequency has played a central role in linguistics. Perhaps the most prominent research in linguistics using frequency was done by Bybee on its role in morpho-phonology and historical analogical changes (Bybee, 1995; Bybee and Hopper, 2001; Bybee, 2001). Generally speaking, token frequency refers to the

number of occurrences of a unit in the language, while type frequency refers to the number of occurrences of a specific pattern. To introduce these concepts more clearly, it is worth using examples from morphology. In morphology, the unit of token and type frequencies is a word. Token frequency is the frequency of a word form, e.g. *broke*. Say *broke* occurred 60 times in a corpus of one million words. Type frequency is the frequency of occurrences of a specific pattern. Say that there are three word forms that have the irregular past tense pattern – *broke, spoke* and *wrote*; the type frequency would then be three. However, it is possible that amongst the word forms of the irregular past tense, their token frequencies differ hugely and therefore each contributes a different amount of weight. Rather than saying that each of the three word forms contributes equally, we would weigh each of them by their respective token frequencies. The type frequency of the irregular past tense is therefore the sum of the log-transformed token frequency of the three word forms. This measure is the weighted type frequency.

Let us return to the level of segments. While there is no doubt that these three measures are highly correlated, they do make different predictions about the nature of segmental frequencies in perceptual confusions. If the two type frequency measures were able to capture more variance than the token frequency measure in perceptual confusions, then one could argue that the listeners are sensitive to lexical information (i.e., the segmental frequencies are computed from the words the listeners know, not from a large sample of segments.). Amongst the two type frequency measures, the literature has conflicted views of whether a weighted measure can better reflect our linguistic knowledge. When calculating neighbourhood density, Bailey and Hahn (2001) proposed a metric that weighs the lexical neighbours of a target word by their respective token frequency. That is, some neighbours contribute more than others. They demonstrated that a weighted measure can capture more variance in behavioural data (non-word acceptability ratings). A later study by Albright

(2007) replicated the analyses in Bailey and Hahn (2001), but the author could not find a significant improvement in the amount of variance explained. In fact, a number of previous studies claimed that pattern strength in the lexicon is determined by type and not token frequency (Bybee, 1995; Albright and Hayes, 2003; Hay, Pierrehumbert, and Beckman, 2004). In sum, the two type frequency measures are expected to outperform the token frequency measure. Furthermore, the weighted type frequency measure is expected to perform worse than the unweighted type frequency measure. If it were to outperform the unweighted measure, the difference should be negligible.

Having described the three frequency measures and their respective predicted performance, we will now briefly describe the two aspects of segmental confusions that are being examined for the existence of any segmental frequency bias. A more detailed description of each of the two aspects can be found in their respective introduction sections.

The first aspect concerns whether frequency can capture the target and response biases. Are certain phones more (or less) likely to be spoken but misperceived (the target)? Are certain phones more (or less) likely to be involved in the resultant perceived phones (the response) in a misperception? If so, whether these patterns can be captured by frequency. In other words, does the frequency distribution of the phones in the language have a similar frequency distribution of the phones being the target and the response of a misperception? From the perspective of a listener, given a segment will be misperceived by the listener, this segment is more likely to be a segment that is frequently spoken in the language, than a segment that is less frequently spoken. Similarly, the responses that the listener gives as the perceived segments (though incorrectly) are biased by the frequency of the segments in the language; that is, the listener would perceive a specific segment more often because this segment is frequent in the language.

The second aspect concerns the asymmetrical confusions. Perceptual confusions are often asymmetrical, i.e. a segment x is perceived as a segment y more often than reverse. We test whether the direction and strength of the asymmetrical confusions across all pairs of segments can be explained by the relative frequency of the two segments in the language. Say that there is an asymmetry between [f] and [θ] in the direction of [θ] > [f]. It is possible that this is due to the fact that [f] is more frequent than [θ] in the language; therefore, listeners are biased to perceive [θ] as [f] more often than the reverse.

In sum, Section 4.2.1 outlines the data that are examined. Section 4.2.2 examines whether frequency can capture the target and response biases. Section 4.2.3 examines whether the strength and direction of the asymmetrical confusions can be predicted by the relative segmental frequencies. Each of the latter two sections contains its own introduction, method, analysis and conclusion sections. Finally, Section 4.2.4 concludes the findings of these two sets of the analyses.

4.2.1 Data extraction

The naturalistic data used in this section are the context-free segmental confusions, as described in Chapter 3, Section 3.2.

Given the three frequency measures, three sets (token, type and weighted type) of actual segmental frequencies were extracted from a control written English corpus as described in Chapter 2, Section 2.3. First, a frequency list was compiled from the corpus. Second, in order to remove words that were erroneously introduced into the corpus due to typos, words that occurred in fewer than three pieces of subtitle texts (i.e. three episodes/films) were removed.[1] Finally, given that we transcribed tapping across word boundaries, words ending in /t/ or /d/ followed by a vowel could have more than one pronunciation (e.g. *it* has two pronunciations: [ɪt] and [ɪɾ]). The

[1] I thank Dr. Emmanuel Keuleers for suggesting this filter.

type frequency measure was computed over the dominant pronunciation, whereas the weighted type frequency measure was computed over the dominant pronunciation but weighted with the combined token frequency of all the pronunciations. The token frequency measure was computed over all the pronunciations and their corresponding token frequencies.

26 consonants were considered – [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, p^h, t^h, k^h, ɾ]. The reason for excluding the glides [j, w] is because in the corpus transcription, they are used as both consonants in onset positions and as offglides of the vowels; therefore, to avoid ambiguity, they were excluded from the consonant set.

14 vowels were considered – [i, ɪ, e, ɛ, æ, a, ɑ, ɔ, o, u, ɜ, ʌ, ʊ, ə], excluding [ʉ] and [ɒ]. The reason for excluding these two vowels is to focus on the General American accent, since the written corpus from which I extract the frequency norms was transcribed with a General American accent, so the frequency norms cannot be found for these two vowels.

4.2.2 Target and response biases

The first aspect of segmental confusions concerns the target and response biases in misperception. Concretely, the questions are whether certain phones are more (or less) likely to be spoken but misperceived (the target), or involved in the resultant perceived phones (the response) in a misperception, and if so, whether this pattern can be captured by frequency. Say that [t] is the most frequently spoken but misperceived segment (i.e. this segment is an intended segment, but it was perceived as something else). Similarly, say that [t] is the most frequently perceived segment for a misperceived intended segment (i.e. a given segment was misperceived as this segment). An obvious explanation would be their high frequencies are simply because [t] is one of the most frequent segments in the language.

It is important to understand that the target bias means that certain phones are more likely to surface as the target of a misperception, and it does *not* mean that certain phones are more likely to undergo misperception. To avoid ambiguity, in this section, we will refer to the segmental frequencies in the language as the *Actual* (segmental) frequencies, the frequencies of an intended segment in a segmental misperception as the *Target* (segmental) frequencies, and the frequencies of a perceived segment in a segmental misperception as the *Response* (segmental) frequencies.

If the actual frequencies correlate with the target frequencies, then it would suggest (given that there is a perceptual error) that the probability of a certain phone being the intended segment of this error is a function of the probability of this phone in the language. In other words, the more frequently an intended segment is produced, the more likely it will be the target of a misperception. This is to say, there is a target bias due to frequency.

Similarly if actual frequencies correlate with response frequencies, then it would suggest that the probability of a given segment being chosen (incorrectly) as the perceived segment is determined by how frequent the perceived segment is in the language. Given an intended segment will be misperceived, the listener will choose a segment as the response based on how frequent it is. This is to say, there is a response bias due to frequency.

In addition to the question of whether there is a target bias and a response bias due to actual frequency, the next question is how much of the variance of these biases can be captured with frequency. The findings from this is crucial, because the variance that cannot be explained by frequency is therefore potentially captured with other non-frequency factors. In terms of the target bias, one non-frequency account is a phonetic account which predicts that a phone that is phonetically less robust (the amount/strength of the phonetic cues) is more likely to be a target of misperception. In terms of the response bias, a non-frequency account is also a phonetic account

which predicts that the choice of the (incorrectly) perceived segment is dependent on its perceived similarity to the intended segment. In sum, this analysis can indirectly highlight the strength of non-frequency factors that are involved in the target and response biases, and the most obvious factor is the phonetic properties of the phones in terms of robustness and their mutual perceived similarity.

These questions were previously examined for naturalistic misperception by Bird (1998) (using 300 instances of naturalistic misperception, which is a sub-corpus of our mega corpus).

Focusing on substitutions, Bird (1998) conducted a correlation analysis separately for consonants and vowels. The author correlated the actual frequencies with the target frequencies, and with the response frequencies. While all the correlations were statistically significant, the correlation values with the vowels were higher (R = 0.89 – 0.93) than those with the consonants (R = 0.80 – 0.84). These high and significant correlations suggest that an extremely high proportion of the variance is explained by the actual frequencies alone. The fact that the correlations were not perfect (i.e. R was not 1) suggests other factors are at work (though playing a very minor role), causing certain phones to be involved in misperceptions more often or less often that the actual frequencies would predict.

Bird's (1998) frequency analyses opened up a range of questions. Firstly, given Bird's (1998) data were based on 300 instances, which is a relatively small sample compared to our mega corpus (around 5,000 instances), can the correlation results be replicated? Secondly, the author focused on substitution. Can we expect to find similar correlations with insertion and deletion?

To conclude, a number of questions can be raised regarding this aspect of segmental confusions. The key question is whether there is a frequency bias for a phone being the target segment and the response segment of a misperception. The second question is, given there is a bias, how strong is this bias? How much of the vari-

ance can be explained with the segmental frequencies alone? The third question is which of three frequency measures (token, type and weighted type) can capture the most variance. The fourth question is whether the findings of the previous questions would differ between consonants and vowels, and between substitutions, insertions and deletions.

4.2.2.1 Method

Given we are interested only in the segments involved in misperceptions, the correctly perceived segments were ignored. That is, the diagonal cells of the confusion matrix were ignored.

A non-parametric correlation, Spearman, was used to compare the two sets of frequencies, since the frequency values are not normally distributed.

4.2.2.2 Analyses

4.2.2.2.1 Consonants The correlation results of the consonants are summarised in Table 4.1. The table contains the correlation values with the level of statistical significance indicated by the number of asterisks. The table categorises the correlation values by the target frequency (substitution and deletion) and response frequency (substitution and insertion) across the table horizontally, as well as by the three frequency measures vertically. The correlation value in bold in each column is the best correlation amongst the three frequency measures.

From the table, we see that all the correlation values are statistically significant and at a strong to very strong level. The lowest value is 0.7820, and the highest value is 0.9670. This clearly indicates that the actual segmental frequency is a strong factor for the target bias and response bias of misperception for consonants.

For both target and response frequencies (substitution, insertion and deletion), type frequency yielded better correlation than the weighted type frequency. This was

	Target		Response	
Frequency Measure	Substitution	Deletion	Substitution	Insertion
Token	0.8417^{***}	$\mathbf{0.9393}^{***}$	0.8273^{***}	$\mathbf{0.9670}^{***}$
Type	$\mathbf{0.9183}^{***}$	0.7824^{***}	$\mathbf{0.9008}^{***}$	0.7936^{***}
Type (Weighted)	0.9042^{***}	0.7820^{***}	0.8943^{***}	0.7841^{***}
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.1: Segmental frequency correlations (Spearman, two-tailed) of consonants between target and response frequencies with actual frequencies of three frequency measures: the superscript symbols denote the level of statistical significance; the bold value in each column is the best correlation amongst the three frequency measures.

expected, since the weighted type frequency is weighted with the token frequency of the relevant lexical items, and previous studies claimed that pattern strength in the lexicon is determined by type frequency and not token frequency (Bybee, 1995; Albright and Hayes, 2003; Hay, Pierrehumbert, and Beckman, 2004). However, it is not always the case that both the type frequency measures (weighted and unweighted) yield better correlations than the token frequency measure. This is only the case for substitution (target and response), while for insertion (response) and deletion (target) token frequency outperforms both measures of type frequency. One possible explanation is that for substitution two lexical items must be involved in the misperception, while for insertion and deletion it is possible (but not necessary) that only one lexical item is involved (i.e. a whole word insertion and a whole word deletion). In this way, insertion and deletion are less sensitive to lexical information than substitution, and yield poorer correlations with the two type frequency measures which are lexically based.

All the correlations are visualised as scatterplots fitted with a linear regression line with confidence intervals. They are Figures 4.1, 4.2, 4.3 and 4.4. Overall, the relative strength of the correlation values is well reflected in the plots, particularly with insertion and deletion.

Although the correlation values are strong, they are not perfect, just as Bird's

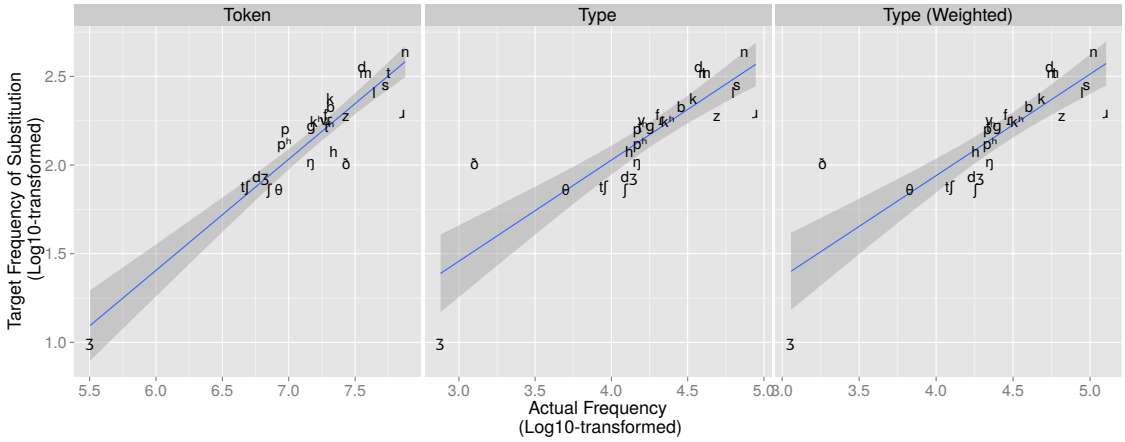


Figure 4.1: The relationship between the target frequencies of substitution and three measures of actual segmental frequencies: consonants

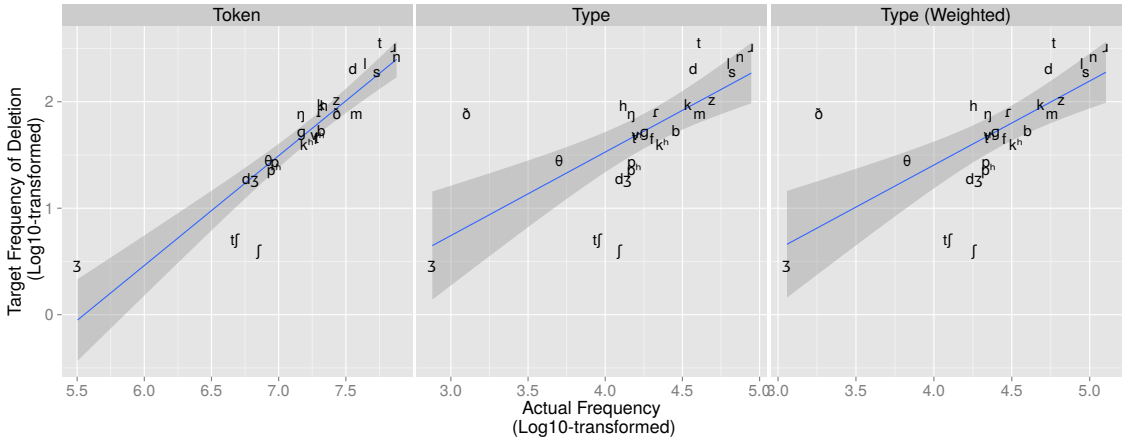


Figure 4.2: The relationship between the target frequencies of deletion and three measures of actual segmental frequencies: consonants

(1998) findings. A closer look at the segments that cannot be explained by the best actual frequency measure (the measure that yields the highest correlation value) could potentially reveal any non-frequency factors. The segments that fall outside the confidence intervals of the linear regression lines (in Figures 4.1, 4.2, 4.3 and 4.4) are summarised in Table 4.2.

First of all, it is worth noting that the diverged segments in the target and response biases of substitution can be analysed together. Similarly, the diverged segments in the target bias of deletion and the response bias of insertion can be analysed

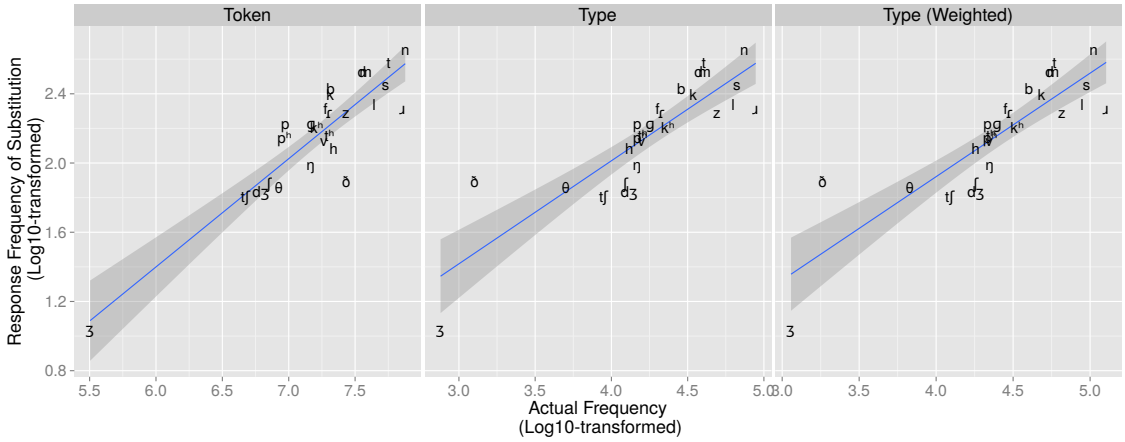


Figure 4.3: The relationship between the response frequencies of substitution and three measures of actual segmental frequencies: consonants

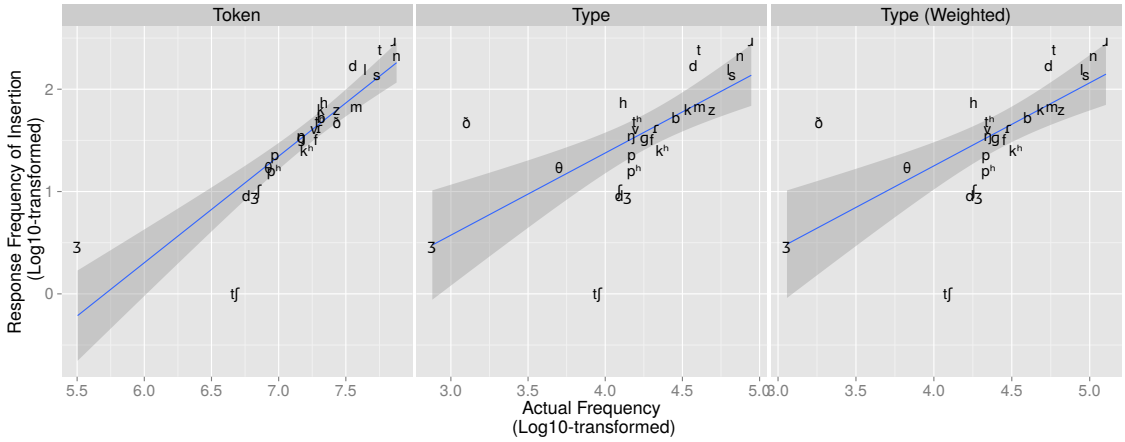


Figure 4.4: The relationship between the response frequencies of insertion and three measures of actual segmental frequencies: consonants

	Target		Response	
	Substitution	Deletion	Substitution	Insertion
More often	[t, d, n, m, ð]	[t, d, ʒ, l, ŋ]	[t, d, n, m, ð, p, b, f]	[t, d, ʒ, l, ɹ, ŋ]
Less often	[z, dʒ, tʃ, ʃ, ʒ, ɹ, ŋ]	[tʃ, ʃ, m]	[z, dʒ, tʃ, ʃ, ʒ, ɹ, l, ŋ]	[dʒ, tʃ, ʃ]

Table 4.2: Consonant segments diverged from actual frequency: the row "More often" denotes the segments that are the target/response of a misperception more often than expected by the best actual frequency measure; the row "Less often" denotes the segments that are the target/response of a misperception less often than expected by the best actual frequency measure.

together. This is because the response patterns can be the result of hypercorrection (Ohala and Shriberg, 1990).

In Table 4.2, the row "More often" contains segments that are the target/response of a misperception more often than expected by the best actual frequency measure (i.e. these segments are above the linear regression). The row "Less often" contains segments that are the target/response of a misperception less often than expected by the best actual frequency measure.

The types of diverged segments are examined in the following order.

1. The segments that are the target and response of a substitution more often than expected
2. The segments that the target of a deletion and the response bias of an insertion more often than expected
3. The segments that are the target and response of a substitution less often than expected.
4. The segments that the target of a deletion and the response bias of an insertion less often than expected

Let us start with the consonant segments that are the target/response of a substitution **more** often than expected. The diverged target segments are [t, d, n, m, ð], and the diverged response segments are [t, d, n, m, ð, p, b, f]. First, [t, d] can be explained by the fact that they are perceptually weak segments (stops are the least sonorous manner) and often undergo lenition intervocalically (Kirchner, 2001). A closer look at the raw confusion matrix in Figure 3.7 in Chapter 3 reveals that [t] and [d] are most confusable with each other, with [t] being perceived as [d] 1.85% of the time, and [d] being perceived as [t] 2.9% of the time. This suggests that the [t] and [d] are diverged from the expected actual frequency due to voicing confusion. Voicing confusion can be viewed as the result of lenition and the hypercorrection of lenition, if the voicing confusions occur intervocalically (further analyses are needed to examine the environment of these voicing confusions).

Second, [n] can be explained by the fact that it occurs **more** often in unstressed environments, e.g. in the word "and" and in prefixes such as "un" and "in". Unstressed environments should be more susceptible to misperception than stressed environments because stressed environments are perceptually more prominent (longer duration, higher intensity). Furthermore, "and", "un" and "in" are highly frequent in the lexicon and they therefore are more likely to have a shorter duration and undergo processes of phonetic reduction (Wright, 1979). Regarding the divergence of [m], a closer look at the raw confusion matrix in Figure 3.7 in Chapter 3 reveals [n] and [m] are most confusable with each other, with [n] being perceived as [m] 2.96% of the time, and [m] being perceived as [n] 5.7% of the time. Given the high confusion between [n] and [m], the divergence of [m] can also be explained. This divergence of [n, m] was also found by Bird (1998).

Third, [ð] can be explained by the fact that it is mainly found in high frequency function words such as "the", "that", "this" "their" etc. Since, high frequency words tend to be phonetically weakened (Wright, 1979), [ð] is misperceived more often than expected by its actual frequency.

Finally, [p, b, f] are the response of a misperception more often than expected by their actual frequencies, for which I have no immediate explanation.

Let us move on to with the consonant segments that are the target of deletion and the response of insertion **more** often than expected. The diverged target segments are [t, d, ʒ, l, ŋ], and the diverged response segments are [t, d, ʒ, l, ɹ, ŋ]. First, [t, d] can be explained by the fact they are often deleted, especially in word-final positions of mono-morphemic words (Guy, 1991; Coetzee and Kawahara, 2013). Second, [l, ɹ] can be explained by the fact that they are often the second/third consonant of a onset cluster (e.g. [pl], [fɹ], [spɹ] etc.). It is well known that these positions are prone to deletion (Harris, 1994), as predicted by their sonority slopes (the deletion should result in a maximal sonority rise) (Ohala, 1999). Finally, I have no explanation for

why [ʒ, ŋ] are inserted or deleted more often than expected.

To briefly conclude, the diverged consonant segments that are the target/response of a substitution/insertion/deletion more often than expected by their actual frequencies can be accounted for using the fact that their phonetic properties are particularly susceptible to misperception.

Next, the consonant segments that are the target/response of a substitution *less* often than expected are examined. The diverged target segments are [z, dʒ, tʃ, ʃ, ʒ, ɹ, ŋ], and the diverged response segments are [z, dʒ, tʃ, ʃ, ʒ, ɹ, l, ŋ]. All the fricatives and affricates [z, dʒ, tʃ, ʃ, ʒ] can be explained by the fact that they are perceptually robust and their acoustic cues lie within the consonants themselves; that is, they are relatively independent of their environment (Wright, 2004). This pattern with the fricatives is consistent with Bird's (1998) findings that the fricatives [s, z] are the target/response of a substitution less often than expected by their actual frequency. Two of the remaining diverged segments are [ɹ, l]. One explanation is that they are liquids which have high acoustic energy, and are high on the sonority scale; therefore, they are particularly salient, and less prone to errors. The last diverged segment is [ŋ] for which I have no explanation since the other nasals have the reverse pattern, [m, n] are the target/response of a substitution *more* often than expected.

Let us move on to with the consonant segments that are the target of deletion and the response of insertion *less* often than expected. The diverged target segments are [tʃ, ʃ, m], and the diverged response segment are [dʒ, tʃ, ʃ]. Similar to the diverged segments with the substitutions, all the fricatives and affricates [dʒ, tʃ, ʃ] can be explained with the fact that they are perceptually robust (Wright, 2004). The remaining diverged segment is [m], for which I have no explanation.

To conclude, most of the diverged consonant segments that are the target/response of a substitution/insertion/deletion more/less often than expected can be explained phonetically. Those that are the target/response more often than expected are pho-

netically weak, while those that are the target/response less often than expected are phonetically strong. Therefore, the target/response patterns that cannot be captured with a frequency account can be captured with a phonetic account.

4.2.2.2.2 Vowels Let us move on to the vowels. The correlation results are summarised in Table 4.3, with the same format as Table 4.1.

	Target		Response	
Frequency Measure	Substitution	Deletion	Substitution	Insertion
Token	**0.8637*****	**0.8185*****	0.8471***	**0.6960****
Type	0.8593***	0.6336*	**0.8845*****	0.5352*
Type (Weighted)	0.8330***	0.6029*	0.8691***	0.5264^{+}
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.3: Segmental frequency correlations (Spearman, two-tailed) of vowels between target and response frequencies with actual frequencies of three frequency measures: the superscript symbols denote the level of statistical significance; the bold value in each column is the best correlation amongst the three frequency measures.

The overall patterns in Table 4.3 are essentially the same as those with the consonants. All but one of the correlations reached statistical significance ($\alpha = 0.05$). The exception is the correlation between the weighted type frequency and insertion.

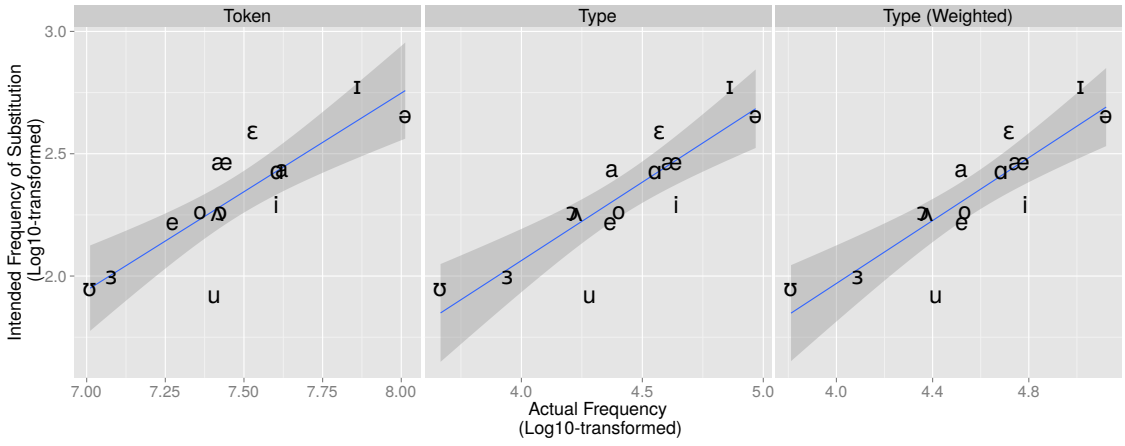


Figure 4.5: The relationship between the target frequencies of substitution and the three measures of actual segmental frequencies: vowels

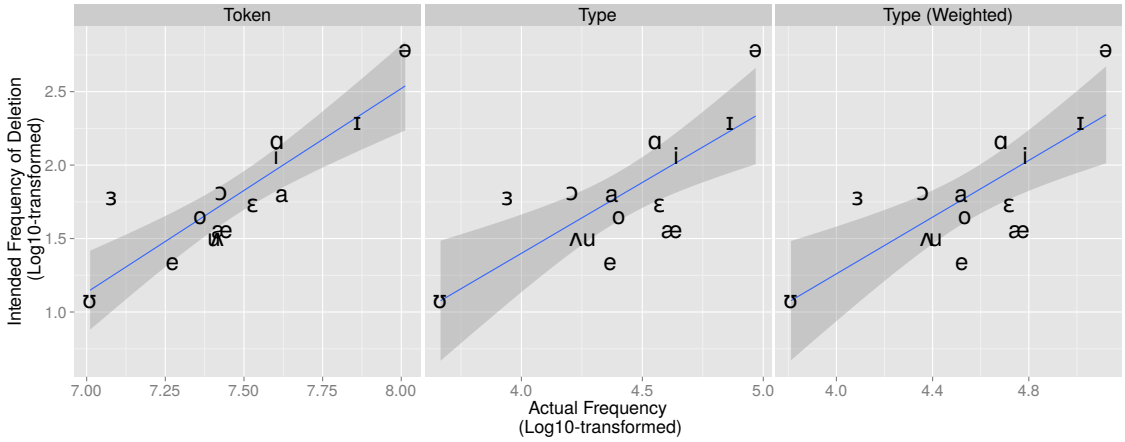


Figure 4.6: The relationship between the target frequencies of deletion and three measures of actual segmental frequencies: vowels

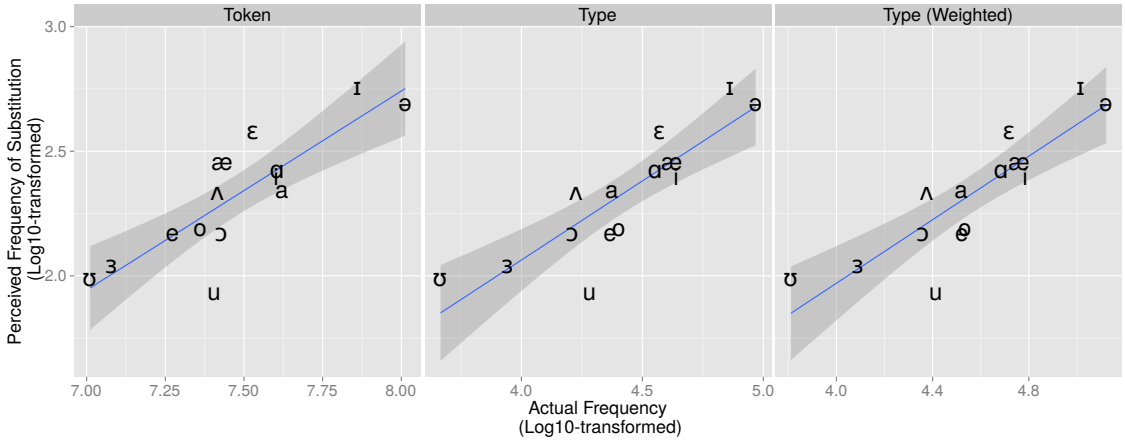


Figure 4.7: The relationship between the response frequencies of substitution and three measures of actual segmental frequencies: vowels

Again, all the correlations are visualised as scatterplots, each fitted with a linear regression line with confidence intervals. They are Figures 4.5, 4.6, 4.7 and 4.8. Overall, the relative strength of the correlation values is well reflected in the plots, particularly with insertion and deletion. Visually, all of the correlations do not appear to be skewed by extreme outliers.

Given the correlation values are not perfect, it is worth examining the diverged segments. The segments that fall outside the confidence intervals of the linear regression lines (in Figures 4.5, 4.6, 4.7 and 4.8) are summarised in Table 4.4. In Table

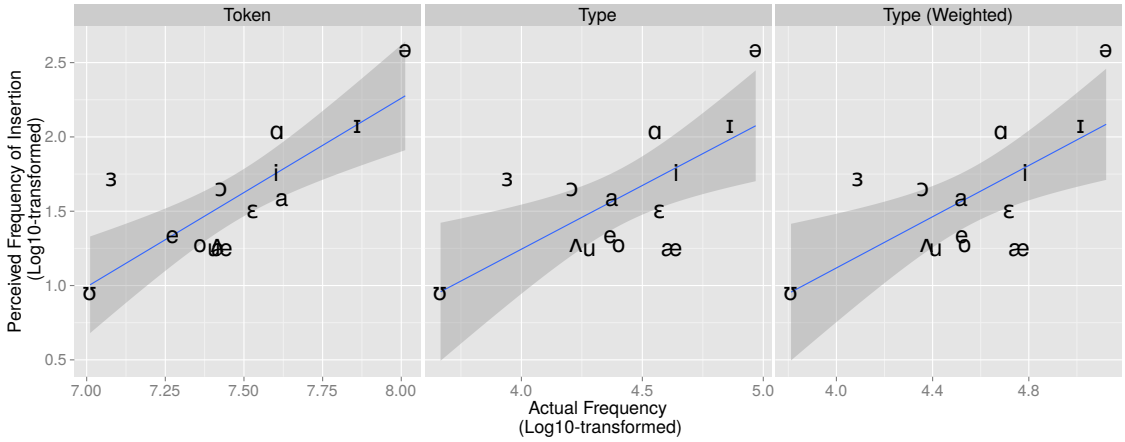


Figure 4.8: The relationship between the response frequencies of insertion and three measures of actual segmental frequencies: vowels

4.4, the row "More often" contains segments that are the target/response of a misperception more often than expected by the best actual frequency measure (i.e. these segments are above the linear regression). The row "Less often" contains segments that are the target/response of a misperception less often than expected by the best actual frequency measure.

	Target		Response	
	Substitution	Deletion	Substitution	Insertion
More often	[æ, ɛ]	[ɜ, ɑ]	[ɪ, ɛ, ʌ]	[ɜ, ɑ]
Less often	[u, i]	[u, a, ʌ, æ]	[u, e, o]	[u, o, ʌ, æ]

Table 4.4: Vowel segments diverged from actual frequency: the row "More often" denotes the segments that are the target/response of a misperception more often than expected by the best actual frequency measure; the row "Less often" denotes the segments that are the target/response of a misperception less often than expected by the best actual frequency measure.

Let us start with the vowel segments that are the target/response of a substitution **more** often than expected. The diverged target segments are [æ, ɛ], and the diverged response segments are [ɪ, ɛ, ʌ]. All the diverged segments are in fact lax vowels, which are phonetically short, with a lower intensity than tense vowels. Therefore, these lax vowels are the target/response of a substitution more often than expected because

they are phonetically weak.

Interestingly, the vowel segments that are the target/response of a substitution **less** often than expected are all tense vowels. The diverged target segments are [u, i], and the diverged response segments are [u, e, o]. Using the same argument as the lax vowels, the divergence of these tense vowels can be explained; tense vowels are longer with a higher intensity than lax vowels. Therefore, they are the target/response of a substitution less often than expected, because they are phonetically strong. Furthermore, these tense vowels in most of the vowel sets transcribed in the naturalistic corpus are followed by an offglide [j] or [w], which makes them more distinctive and less prone to misperception.

The vowel segments that are the target of deletion and the response of insertion more often than expected are [ɜ, ɑ]. The divergence of [ɜ] can be explained with its high confusion with [ə] and [ɑ]. [ɜ] is mostly often perceived as [ɑ] 2.45% of the time, followed by [ə] 2.31% of the time (see Figure 3.8 in Chapter 3). [ɜ] being confused as [ə] is perhaps due to their close acoustic distance. However, I have no explanation for why [ɜ] is perceived most often as [ɑ].

The vowel segments that are the target of deletion and the response of insertion less often than expected are [u, o, a, ʌ, æ]. I have no explanation for these diverged segments, since they are a mixture of tense and lax vowels, front and back vowels, and close and open vowels. To conclude, just as the diverged consonant segments, most of the diverged vowel segments can be explained using a phonetic account.

Finally, the results of the consonants and the vowels are compared. For the vowels, the correlation value ranges from 0.5264 to 0.8845, while tor the consonants, the correlation value ranges from 0.7820 to 0.9670. While the correlation values of the vowels were all relatively high, they are lower than those of the consonants. This indicates that the amount of frequency bias is higher for the consonants than for the vowels. The opposite is true with the phonetic bias, as found in Section 3.5 and

Section 3.6 in Chapter 3. The amount of phonetic bias was stronger for the vowels than for the consonants. Furthermore, the analysis of the diverged consonants and vowels suggests that what cannot be captured with a frequency bias can be captured with a phonetic bias. Together, one could speculate that the amount of frequency bias complements the amount of phonetic bias in segmental confusions; that is to say, they have an inverse relationship.

Regarding the three frequency measures, again unweighted type frequency outperforms weighted type frequency. The two measures of type frequency outperformed token frequency for substitution, but only for the response frequencies. Just as with consonants, we found that token frequency outperforms both measures of type frequency for insertion and deletion. The earlier discussion of the three frequency measures for the consonants also applies to the vowels.

4.2.2.3 Conclusion

This section examined whether the target and response frequencies in segmental misperception can be explained using the actual frequencies in the language. This question was examined for substitution (target and response), insertion (response) and deletion (target) errors of consonants and vowels.

Section 4.2.2.2.1 examined the substitution, insertion and deletion errors of consonants. The strength of the correlations was at a strong to very strong level (ρ = 0.7820 – 0.9670). Section 4.2.2.2.2 examined the substitution, insertion and deletion errors of vowels. Again, the strength of the correlations was strong (ρ = 0.5264 – 0.8845). These strong correlations indicate that the actual segmental frequencies in the language are a strong factor for the probability of a certain phone being a target or a response of a misperception. To recap, we are *not* referring to the probability of a phone being misperceived, and we are referring to the probability of a phone being a target or a response of a misperception, given there is a misperception.

Concretely, the segment [n] is a more frequent segment than [ʒ]. Given a phone x will be misperceived, this phone x is more likely to be [n] rather than [ʒ]; therefore, the target is biased by the actual frequency. Similarly, given a phone x will be misperceived, the perceived phone y is more likely to be [n] rather than [ʒ].

Most of the segments that diverged from their actual frequencies can be explained using a phonetic account. With the consonants, the segments that were the target/response of a substitution/insertion/deletion more often than expected by their actual frequencies were 1) phonetically weak – [t, d], 2) susceptible to cluster reduction – [l, ɹ]) and 3) susceptible to phonetic weakening due to lexical frequencies – [n] in "and", and [ð] in "the". Similarly, the consonant segments that were the target/response of a substitution/insertion/deletion less often than expected by their actual frequencies were mostly fricatives and affricates which are phonetically strong. With the vowels, there was a clear tense-lax difference with substitutions. Lax vowels were the target/response of a substitution more often than expected by their actual frequencies, because lax vowels are phonetically weak. Tense vowels were the target/response of a substitution less often than expected by their actual frequencies, because tense vowels are phonetically strong.

Furthermore, we found that the correlation values are lower for vowels than for consonants, which indicates that consonants are more sensitive to this frequency bias than vowels. Given that the opposite is true with phonetic bias, as found in Chapter 3 that the diverged segments can mostly be explained using a phonetic account, I speculated that the amount of frequency bias has an inverse relationship with the amount of phonetic bias in segmental confusions. Further analyses are needed to substantiate this speculation by regressing (e.g. with a regression model) the confusion patterns with *both* the frequency bias and the phonetic bias, because it is possible that the phonetic bias also correlates with the frequency bias.

Again, it was found that two measures of type frequency outperformed token

frequency for substitution. This advantage of type frequency will also be found in Section 4.4.2. Surprisingly, token frequency outperformed both measures of type frequency for insertion and deletion. Given that type frequencies are lexically-based, one explanation is that this difference between substitution and insertion/deletion is due to the fact that insertion and deletion are less sensitive to lexical information than substitution, because substitution errors have to involve two lexical items, while insertion and deletion errors could involve only one (i.e. whole word deletions/insertions).

4.2.3 Asymmetrical confusion

The second aspect of segmental confusions concerns their asymmetrical patterns. Recall that three asymmetrical patterns (namely TH-fronting, velar nasal fronting, and back vowel fronting) in naturalistic and experimental misperception were analysed in Chapter 3, Section 3.8. Indeed, all three patterns were confirmed, with [θ] being perceived as [f], [ŋ] as [n] and back vowels as front vowels, more often than the reverse. Finally, we used them as evidence for a perceptual-based account of sound change. However, it is possible that their asymmetries are affected by their relative segmental frequencies. For instance, say that [f] is more frequent in the language than [θ]; [f] could then be chosen as the perceived segment for the intended segment [θ] more often than the reverse, because there is a response bias due to frequency differences. This asymmetrical pattern of [θ] > [f] can therefore be explained without the need of invoking accounts of perceptual biases.

It is worth noting that this bias is similar to the response bias mentioned earlier in Section 4.2.2. Nonetheless, they differ in terms of whether the correctly perceived segments are considered. The bias in Section 4.2.2 concerns only the segments that are involved in a segmental misperception and not the correctly perceived segments, while the current bias concerns both because asymmetricality depends on the propor-

tions of correctly and incorrectly perceived segments (see Chapter 3, Section 3.8.1 for the method for calculating asymmetries).

Benkí (2003) conducted an analysis of whether the asymmetrical patterns in segmental confusions can be captured under a frequency/lexical account. Using experimentally induced misperception of nonsense CVC syllables, the author computed the strength and direction of the asymmetries using the criterion measure (henceforth c bias) from choice theory of eleven pairs of segments. These eleven pairs of segments were three onset pairs – [t, p], [k, p] and [ɹ, l], three vowel pairs – [æ, ɑː], [u, i] and [o, e], and five coda pairs – [t, p], [k, p], [k, t], [g, d] and [m, n]. The relative frequency measures were computed by subtracting the frequency of one of the two segments in a given pair from the frequency of the other segment in the same pair. Four different frequency measures were tested separately. They are a) the number of occurrences per 100 phonemes, b) the number of lexical items containing the phoneme, c) the number of occurrences per million words, and d) the sum of the log-transformed frequencies of the lexical items containing the phoneme. In fact, c) is virtually the same as our token frequency measure, and b) and d) are the same as our two measures of type frequency. The author found that on the whole all of the frequency measures captured a sizable portion of the variance (R^2 from 0.2 to 0.3) of the c bias values; however, none of them were statistically significant at $\alpha = 0.05$. Of the four frequency measures, the number of lexical items containing the phoneme (type frequency) captured most variance, $R^2 = 0.290$, and with the smallest p-value, $p = 0.088$, which is near-significant.

Benkí's (2003) findings are encouraging. The high level of variance explained across multiple relative frequencies indicates that the relative frequency of the two segments can predict the strength and direction of their confusions. Although the p-values did not reach significance, it is likely that this is due to the small number of pairs tested (11 pairs). Therefore, by testing more segmental pairs, we could then

have a more complete picture of whether the relative frequency is a useful factor for predicting asymmetries. In the current analyses, all segmental pairs are tested separately for consonants and vowels. Furthermore, just as Benkí (2003), multiple frequency measures are examined.

To conclude, a number of questions can be raised. Firstly, can the strength and direction of the asymmetrical pattern for each pair of phones be captured by the relative segmental frequencies in the language? Secondly, how much variance can be captured? Thirdly, which of the three frequency measures can capture the most variance? Finally, would the findings of the previous questions differ between consonants and vowels?

4.2.3.1 Method

Regarding the consonant pairs, with 26 consonants, 325 consonant pairs are possible. For the vowel pairs, with 14 vowels, 91 vowel pairs are possible. The strength and direction of the asymmetries were estimated using the criterion measure (c bias) as described in Chapter 3, Section 3.8.1. The c bias values were computed for all 325 consonant pairs and all 91 vowel pairs. Just as in Chapter 3, Section 3.8.1, we excluded pairs that have no confusion in either direction, because their resultant c bias values are dependent purely on the smoothing process. The order of the two segments in a given pair can affect the sign of the c bias; therefore, it is worth establishing a notation system for later reference. For a given pair [Segment 1 > Segment 2], a positive c bias value means that Segment 1 is perceived as Segment 2 more often than the reverse, a negative c bias value means the Segment 2 is perceived as Segment 1 more often than the reverse, and a zero c bias value means that there is no asymmetrical confusion. The first segment in a given pair is referred to as Segment 1 and the second segment as Segment 2.

The relative frequency of each segmental pair was calculated by taking a ratio of

the frequency of Segment 2 and the frequency of Segment 1. To remove the skewness of frequency values, the ratios were then log-transformed for all three measures. This is summarised as the following metric: $Log10(Frequency_{Seg2}/Frequency_{Seg1})$. A positive log-ratio means that Segment 2 is more frequent than Segment 1, a negative log-ratio means that Segment 1 is more frequent than Segment 2, and a zero log-ratio value means that the two segments are equally frequent. This log-ratio has a further advantage of having zero as the centre of the scale just as the c bias value. Therefore, if relative frequencies can predict asymmetries, then a positive correlation is expected between the log-ratios and the c bias values.

A non-parametric correlation, Spearman, was used to compare the two sets of frequencies, since the data are not normally distributed; therefore, a non-parametric correlation is more appropriate.

4.2.3.2 Analyses

Table 4.5 summarises the correlation analyses for consonants and vowels between the c bias values (which reflect the confusion asymmetries) and the log-ratios (which reflect the frequency asymmetries). The table shows the correlation values (Spearman, two-tailed) as well the level of statistical significance.

Frequency Measure	Consonants	Vowels
Token	**0.8068*****	**0.8478*****
Type	0.7080***	0.7851***
Type (Weighted)	0.7109***	0.7847***
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$		

Table 4.5: Correlations (Spearman, two-tailed) between confusion asymmetries and frequency asymmetries of consonants and vowels with three frequency measures: the superscript symbols denote the level of statistical significance; the bold value in each column is the best correlation amongst the three frequency measures.

All the correlation values are highly significant at a strong to very strong level (ρ = 0.71 – 0.85). Token frequency yields higher correlation values than the two

measures of type frequency. The unweighted frequency measure does not outperform the weighted one consistently, only with vowel asymmetries, and not with consonant asymmetries. Finally, we see that the correlations with the vowels are stronger than those with the consonants. These findings are surprising, considering in the previous analyses of segmental frequency in Section 4.2.2 we found the exact opposite patterns – a) consonants are more affected by frequency than vowels and b) the two measures of type frequency outperform token frequency. Regarding how the vowels are more affected by frequency than consonants in terms of confusion asymmetries, I have no immediate explanation. Regarding the sudden advantage of token frequency in predicting confusion asymmetries, one explanation lies in how asymmetries are defined. Recall in Chapter 3, Section 3.8.1, we described the criterion measure (c bias) which is used to reflect the confusion asymmetries. The c bias measure relies on the proportion (not count) of confusions in each direction, and the frequencies of the correctly perceived segments (the diagonal cells in a confusion matrix) are required to compute the proportions. In the naturalistic corpus, the frequency of the correctly perceived segments should highly correlate with their frequency in the language, because the naturalistic corpus is a sample of the language. This is indeed the case, as indicated by the correlation (Spearman, two-tailed) between the frequency of the correctly perceived segments and their frequency in the language. With the consonants, the correlation values are 0.9835, 0.8427 and 0.8345, with token, type and weighted type frequency respectively, and they were all highly significant. With the vowels, the correlation values are 0.9648, 0.8373 and 0.8109, with token, type and weighted type frequency respectively and they were all highly significant. Given that the correctly perceived segments are extracted from all the segments in all words that are correctly perceived in the corpus, these extracted frequencies are, in fact, token frequency, and not type frequency. Therefore, the advantage of token frequency in predicting the confusion asymmetries can be explained.

To examine these relationships further, we will visualise the correlations as scatterplots, fitted with linear regression lines. The scatterplots of the consonants are shown in Figure 4.9 and those of the vowels are shown in Figure 4.10. Focusing on the consonants, Figure 4.9 shows that the regression lines with the two measures of type frequency are poorly fitted, compared to that with token frequency. A closer inspection of the segmental pairs reveals that the poor fits are due to [ð] having a low type frequency, as [ð] is found in all the pairs that are outliers (visually). Let us move on to the vowels. Figure 4.10 shows that the regression line has a tighter fit with token frequency than with the two measures of type frequency. However, unlike the consonants, visually the differences cannot be attributed to specific segments.

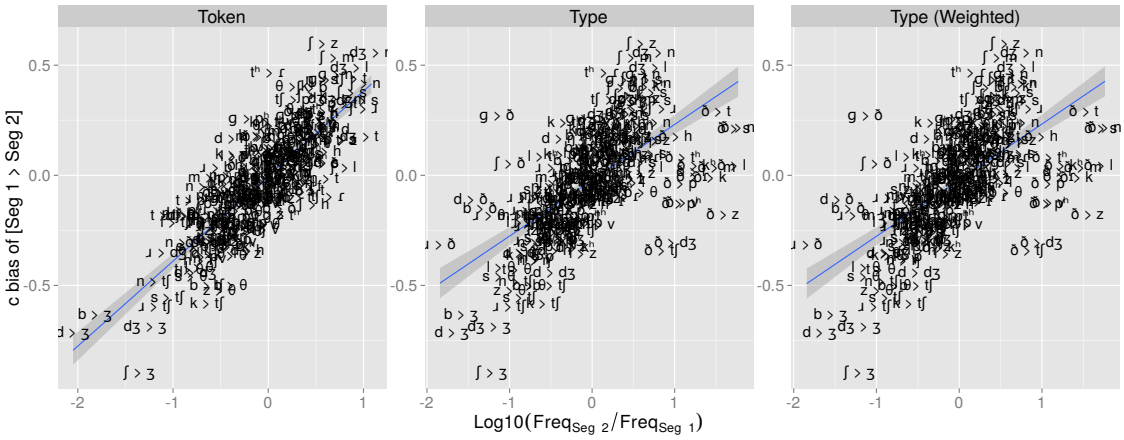


Figure 4.9: The relationship between confusion asymmetries and frequency asymmetries of consonants

Finally, based on our results in this section, we should reconsider our conclusion based on the analyses in Chapter 3, Section 3.8, where we analysed the three asymmetrical patterns, namely TH-fronting, velar nasal fronting, and back vowel fronting. Our results in this section suggest that confusion asymmetries are affected by the relative segmental frequencies found in the language, by examining all the possible asymmetries (i.e. all combinations of two segments). In fact, TH-fronting and velar nasal fronting can both be explained using a frequency account, since [n] is more frequent than [ŋ], and [f] is more frequent than [θ].

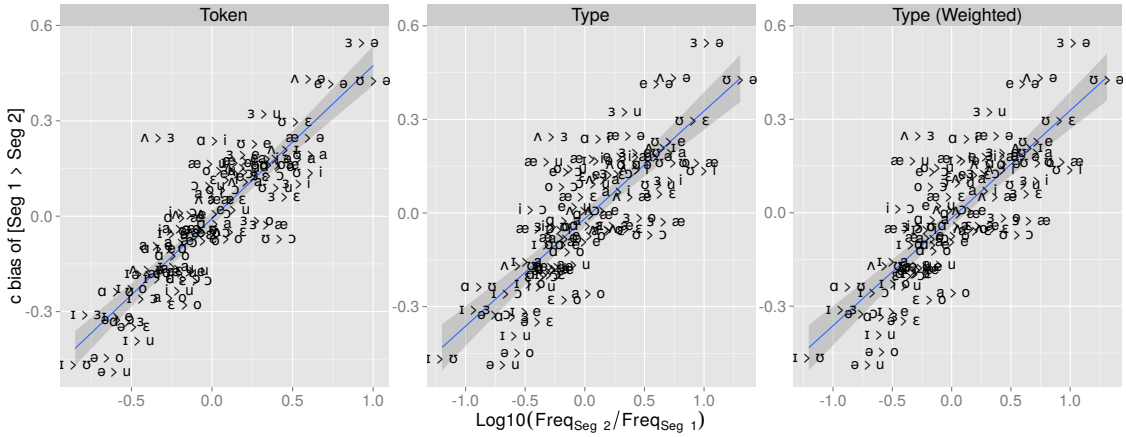
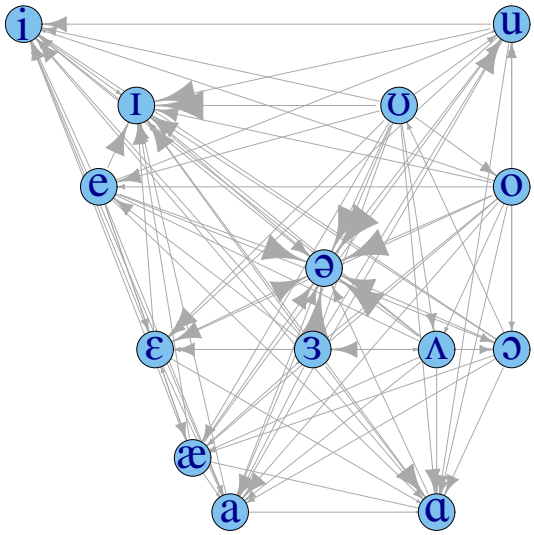


Figure 4.10: The relationship between confusion asymmetries and frequency asymmetries of vowels

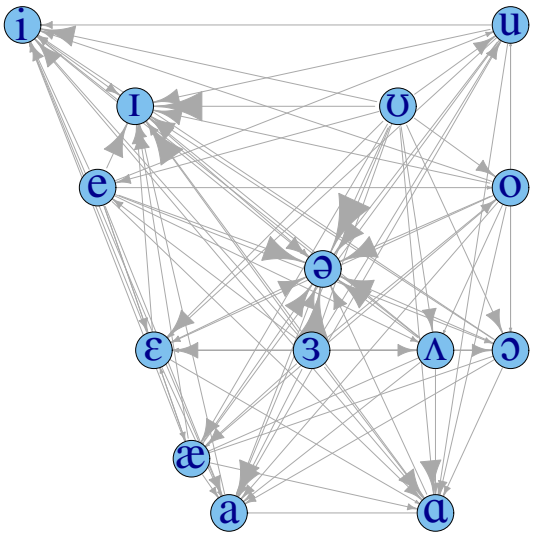
By visualising the direction and strength of the asymmetries with a vowel chart, we can better evaluate the back vowel fronting pattern. The asymmetrical patterns for all the vowel combinations are shown in Figure 4.11. Figure 4.11 contains three sub-figures. Figure 4.11a summarises the confusion asymmetries. Figure 4.11b summarises the token frequency asymmetries. Figure 4.11c summarises the type frequency asymmetries (there are no visual differences between the weighted and unweighted type frequency measures). In each figure, each vowel is connected with all other vowels with a straight line, the direction of the arrow head reflects the direction of the asymmetry, and the size of the arrow head reflects the strength of the asymmetry. Visually, all three figures are extremely similar in terms of the direction of the arrows.

A back vowel fronting pattern can indeed be found in the confusion asymmetries (Figure 4.11a) as all the back vowels (except [ɑ]) have fewer incoming arrows than the front vowels. However, the same pattern can be found also with the token frequency asymmetries and type frequency asymmetries in Figure 4.11b and Figure 4.11c respectively. This suggests that the back vowel fronting pattern is affected by segmental frequencies.

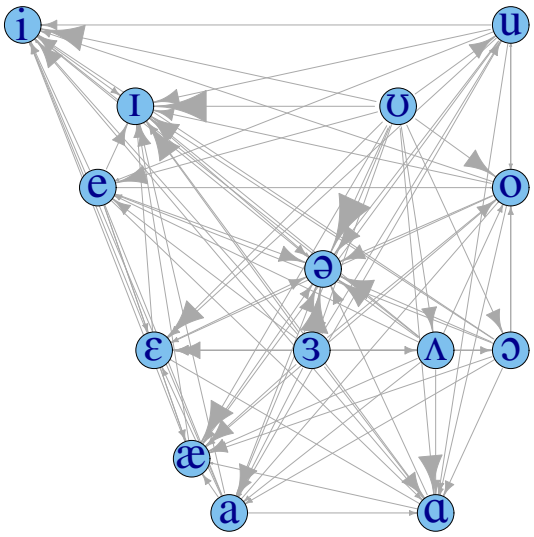
Although unrelated to the fronting pattern, it is worth noting that there is a



(a) Confusion asymmetries



(b) Token frequency asymmetries



(c) Type frequency asymmetries

Figure 4.11: Visualisation of vowel asymmetries: a) confusion asymmetries, b) token frequency asymmetries, c) type frequency asymmetries (both weighted and unweighted).

centring pattern with most vowels moving into [ə] in both confusion asymmetries and frequency asymmetries. This centring pattern in frequency asymmetries could in fact explain an earlier observation in Chapter 3, Section 3.4.3.1, that there is a confusion bias of open/close vowels being perceived as mid vowels, which is essentially

centring.

4.2.3.3 Conclusion

This section examined the effect of frequency on the asymmetrical patterns in segmental confusions. In Chapter 3, Section 3.8, we analysed three asymmetrical patterns, namely TH-fronting, velar nasal fronting, and back vowel fronting, in naturalistic and experimental misperception. As an extension to the previous analyses, we correlated the frequency asymmetries (the difference in frequency between two segments) and the confusion asymmetries of all possible pairs of segments for both vowels and consonants. The correlations were highly significant at a strong to very strong level ($\rho = 0.71 - 0.85$). This is a surprising finding, suggesting that confusion asymmetries can be affected by frequency asymmetries.

Indeed, the three asymmetrical patterns, TH-fronting, velar nasal fronting, and back vowel fronting, were explicable using frequency, since [n] is more frequent than [ŋ], and [f] is more frequent than [θ], and front vowels are generally more frequent than back vowels. These results suggest that we should reconsider our conclusion in Chapter 3, Section 3.8, such that confusion symmetries are a function of not only perceptual biases (Ohala, 1981; Ohala, 1989), but also frequency biases.

Furthermore, we found that vowels are more affected by frequency than consonants and that token frequency outperformed type frequency. These are the exact opposite patterns from those found in Section 4.2.2. I have no explanation for why vowels are more affected by frequency than consonants. However, regarding the advantage of token frequency, I proposed that it is due to how confusion asymmetries are defined. The criterion measure (c bias) which is used to reflect the confusion asymmetries relies on the frequencies of the correctly perceived segments (the diagonal cells in a confusion matrix). The correctly perceived segments are extracted from all the segments in all words that are correctly perceived in the corpus. Therefore,

they are token frequency, and not type frequency. The advantage of token frequency should not be interpreted as a linguistic advantage, but as a methodological bias.

4.2.4 Conclusion

This section conducted separate sets of analyses to examine the role of segmental frequencies in segmental confusions. We focused on two aspects of segmental confusions that could be the result of the segmental frequencies in the language, namely a) the effect of frequency on target and response biases, and b) the relationship between frequency asymmetries and confusion asymmetries.

Section 4.2.2 found that the target and response biases can be captured nearly perfectly with segmental frequencies as indicated by the strong to very strong level of correlations. This pattern is true for substitutions, insertions and deletions. We found that the effect of frequency is stronger for consonants than for vowels.

Section 4.2.3 found that confusion asymmetries can be affected by frequency asymmetries, as indicated by the strong level of correlations. The theoretical implication of this is that confusion asymmetries are modulated by top-down factors such as segmental frequencies. In other words, the result suggests that confusion symmetries is a function of both perceptual biases (Ohala, 1981; Ohala, 1989), and frequency biases. Further work is needed to examine whether frequency asymmetries play a role in experimental misperception such as Miller and Nicely (1955).

In terms of the comparison between the different frequency measures, overall we found type frequency outperformed token frequency in substitution (but not insertion and deletion), which supports the claims that pattern strength in the lexicon is determined by type, and not token frequency (Bybee, 1995; Albright and Hayes, 2003; Hay, Pierrehumbert, and Beckman, 2004).

Together all these findings allow us to make the following conclusions. Given that a segment will be misperceived, the target and the response are determined

by the actual frequencies; that is, more frequent segments are more likely to be misheard (target) and incorrectly chosen as the perceived segment (response). This holds for substitution, insertion and deletion errors. Crucially this frequency bias operates independently on the intended segments and the perceived segments. Since frequency can partly determine the intended and perceived segment, it can also bias the overall confusion patterns, namely the asymmetricality of confusions.

To conclude, these findings confirm the fact that listeners are sensitive to frequency information on a segmental level in misperception, and as such there is a top-down effect from the lexicon.

4.3 Syllable factors

The previous section examined the role of segmental frequency in segmental misperception. Let us move away from the segment level factors. This section will examine whether factors on the syllable level play a role in segmental misperception.

We will focus on the rate of segmental errors, ignoring the nature of the perceived segments (what the intended segments are misperceived as). Three syllable-based factors are examined for their effect on the error rates. The first factor is the *syllable constituency* – the position of the segment in a syllable (namely onset, nucleus and coda). Do segmental errors occur evenly across the three constituents? The second factor is the *syllable position* – the position of the syllable (that contains the segment) in a word. For a polysyllabic word, three positions can be generalised, namely word initial, word medial, and word final. For a monosyllabic word, this three way categorisation cannot be applied. Do segmental errors occur more often in word-final syllables than word initial syllables? The third factor is *stress* – whether the segment is in a stressed or unstressed syllable. Do segmental errors occur more in unstressed than stressed syllables? These factors (apart from syllable position which

is only relevant in the polysyllabic words) are examined separately for monosyllabic and polysyllabic words.

4.3.1 Syllable constituency

The three constituents are onset, nucleus and coda. Using phonetic arguments and previous experiment findings, predictions can be made as to which constituents are more likely to be misperceived.

Firstly, the difference between the nucleus, and the onset/coda is apparent, as the nucleus contains vowels and the onset/coda contains consonants. Vowels are often longer, and acoustically more intense (cf. sonority) than consonants. Recall that in Section 3.4.1 of Chapter 3 we analysed the overall error rate of vowels and consonants, and we found that consonants are more erroneous than vowels with the rates 19.96% and 17.67% respectively. Therefore, we would naturally expect that the nucleus is less erroneous than the onset/coda.

Secondly, the perceptual difference between onset and coda is less clear. On the one hand, onsets are argued to have a higher degree of cue redundancy, e.g. there is greater redundancy of cues in the CV transition than VC transitions, which is especially true in stop consonants which have VOT and always have release bursts (Wright, 2004). On the other hand, codas are predictable by their preceding nucleus. For instance, the length of the vowel can cue the voicing of the final consonants (pre-fortis clipping) and vowels have been shown to lengthen before fricatives (Peterson and Lehiste, 1960). Vowel nasalisation is mainly caused by nasal codas, so the presence of vowel nasalisation serves as a stronger cue for nasal codas than for nasal onsets. Furthermore, there are studies that suggest the cues of codas are spanned over a greater duration than those of onsets. For instance, formant two and formant three have greater movement in codas than in onsets (Broad and Fertig, 1970); and the transition durations tend to be longer in VC than CV positions (Lehiste and

Peterson, 1961). Besides the acoustic information, the phonotactic information of English makes codas more predictable, because the range of codas is more restricted than that of onsets (Kessler and Treiman, 1997).

Thirdly, while we would naturally expect that the nucleus is less erroneous than onset/coda because of the difference between vowels and consonants, it is possible that the nucleus is just as erroneous as the coda, with the onset being the most erroneous – Onset > Nucleus/Coda (">" means more erroneous than). This prediction is supported by the following facts. Firstly, the phonotactic analyses of Kessler and Treiman (1997) found that consonants have a different distribution within the rime than outside the rime. That is, the co-occurrence constraints lie with the nucleus and the coda more than with the onset and the nucleus in a CVC syllable. Secondly, the phonetic arguments given in the previous paragraph do not only highlight the perceptual salience of codas, but also the fact that the nucleus and the coda overlap more in terms of their acoustic cues than the nucleus and the onset.

Besides using phonetic arguments to form our predictions, we could review previous confusion experiments which tested the error rates of these constituents. The classic confusion study by Wang and Bilger (1973) tested the confusiability of consonants in both CV and VC syllables. As summarised in Chapter 3, Section 3.7.1.3, Wang and Bilger (1973) tested two sets of CV and VC syllables composed of 24 consonants and three vowels. The first set of CV and VC (CV-1 and VC-1) contains the same set of consonants [p], [t], [k], [b], [d], [g], [f], [θ], [s], [ʃ], [v], [ð], [z], [ʒ], [tʃ], [dʒ]. The second set of CV and VC (CV-2 and VC-2) contains different consonants with [p], [t], [tʃ], [dʒ], [l], [ɹ], [f], [s], [v], [m], [n], [h], [hʷ], [w], [j] for CV-2, and [p], [t], [g], [ŋ], [m], [n], [f], [θ], [s], [ʃ], [v], [ð], [z], [ʒ], [tʃ], [dʒ] for VC-2. On the one hand, CV-2 and VC-2 are in a sense more realistic, as they contain consonants that can only be in either CV or VC, namely [h] and [ŋ]. On the other hand, CV-1 and VC-1 are balanced, which allows for a more direct comparison of onsets and

codas. Codas were more erroneous than onsets in the second syllable set (CV-2 and VC-2); however, this difference is inconsistent with the first syllable set, with codas being more erroneous only at more difficult signal to noise ratios and with the vowel [ɑː, and uː], but not [iː] (Wang and Bilger, 1973, pp. 1251–1252). Despite the inconsistency with CV-1 and CV-2, codas were, overall, more erroneous than onsets. In another confusion experiment, Cutler et al. (2004) tested all possible standard American English CV and VC sequences. However, unlike Wang and Bilger (1973), onsets were more erroneous than codas (with an average 5% difference in error rate).

Besides the confusion experiments of CV, VC syllables, Redford and Diehl (1999) tested 147 CVC syllables (7 consonants × 3 vowels × 7 consonants), and found that codas are more erroneous than onsets. They conducted a further acoustic analysis of the stimuli and found that the perceptual advantage of onsets is partly due to their longer duration and higher amplitude. That is, onsets are produced with greater acoustic distinctiveness than codas. In another CVC confusion experiment conducted by Benkí (2003), it was found that codas were more erroneous than both onsets and nuclei, and that onsets and nuclei are similarly erroneous (Benkí, 2003, pp. 137–140).

The conflicting findings between Cutler et al. (2004) and the other studies (Redford and Diehl, 1999; Benkí, 2003) were examined by Cutler et al. (2004). The author subsetted their data to best match the consonants and vowels used in Redford and Diehl (1999) and Benkí (2003). After the subsetting, they still found onsets being more erroneous than codas. The conflicting findings are therefore likely due to the different experimental conditions, such as the signal to noise ratio, the number of consonants and vowels tested, the number of speakers and listeners tested, etc. Given this mismatch between Cutler et al. (2004) and the other experimental studies, the confusion patterns in the naturalistic corpus could, in fact, be used to settle the debate. In any case, there is converging evidence from multiple confusion studies

(Wang and Bilger, 1973; Redford and Diehl, 1999; Benkí, 2003) with the exception of Cutler et al. (2004) that codas are more erroneous than onsets. In addition, Benkí (2003) found that onsets and nuclei are similarly erroneous.

In sum, considering both phonetic arguments and the relative error rates found in previous experimental confusion data, it is unclear what the general pattern is. In fact, a range of predictions can be made regarding the relative error rates of onset, nucleus and coda. They are summarised below (">" means more erroneous than).

- *[Onset, Coda] > Nucleus*
- *Onset > Coda > Nucleus*
- *Coda > Onset > Nucleus*
- *Coda > [Onset, Nucleus]*
- *Onset > [Nucleus, Coda]*

Furthermore, it is unclear whether these predictions would hold for both monosyllabic and polysyllabic words, given that all the above arguments were based on data on single syllables. The perceptual/phonetic arguments were based mostly on experimental data that tested only single syllables (Peterson and Lehiste, 1960; Wright, 2004). The same is true for the phonotactic analysis by Kessler and Treiman (1997) which was also based on single syllables (monomorphemic CVC words). Finally, the experimental confusion data are also restricted to single syllables (CV, VC or CVC).

To conclude, amongst these predictions, the most likely one is the one based on experimental confusion studies, *Coda > [Onset, Nucleus]*, because we are also analysing confusion data (though naturalistic, not experimental). Since the experimental data were based on single syllables, the most conservative prediction is that the error rates in monosyllabic words have the trend – *Coda > [Onset, Nucleus]*, with coda being more erroneous than onset/nucleus. We lack specific predictions for polysyllabic words.

4.3.2 Syllable position

In order to form a prediction regarding the effect of syllable position on error rates, we first consider the acoustic realisation of the segments in word-initial, medial and final positions.

In an acoustic analysis, Lindblom (1968) tested the effect of syllable position on the duration of segments. The author found that segments are longer in final syllables than in medial syllables, which in turn are longer than the segments in initial syllables. This lengthening effect holds for both unstressed and stressed syllables and can be attributed to a final lengthening effect.

The final lengthening effect (Klatt, 1975) is the effect of lengthening the final word of a phrase. Traditionally, experimental studies have examined only the final syllable of the final word, but in fact there is evidence in American English that the lengthening begins before the final syllable (Turk and Shattuck-Hufnagel, 2007). In a production study of American English by Turk and Shattuck-Hufnagel (2007), they found that, besides the final syllable (especially the rime), other regions are lengthened as well. These are the rime of the main stressed syllable (when it is not word final), and the regions between the main stressed syllable and the final syllable (though only sporadically). In sum, generally there is a progressive lengthening effect across the final word of a phrase with an increase of lengthening towards the final syllable.

Using the patterns found in the final lengthening effect, assuming that most words can appear as the final word of a phrase, word final syllables can therefore on average be longer than word medial syllables, which can be in turn longer than word initial syllables. As such, this can be used to form a prediction that segments in word final syllables are less erroneous than those in word medial syllables and in turn are less erroneous than those in word initial syllables, because the longer the duration of a segment, the more perceptually salient it is, which means it is less erroneous. This

can be summarised as: *Word Initial* > *Word Medial* > *Word Final* (">" means more erroneous than).

Alternatively we could extend the phonotactic account of predictability mentioned in Section 4.3.1. Concretely, the number of possible segments in a given segment position in a word decreases as the position moves from left to right. That is, the number of lexical candidates decreases with every additional segment perceived. This is essentially the idea of uniqueness points in word recognition (Luce, 1986a). This could mean that segments in word-final syllables are more predictable than those in earlier syllables, because there are fewer potential lexical candidates. These predictable segments will therefore have lower error rates. In sum, this predictability account makes the same prediction as the duration account, with the same trend *Word Initial* > *Word Medial* > *Word Final*.

4.3.3 Stress

Acoustically speaking, stressed syllables are more perceptually salient than unstressed syllables, such that they have longer duration, higher intensity and they can carry extreme intonation (Browman, 1980). Indeed, a stressed syllable has been argued to be an "island of reliability" (Pisoni, 1981). That is, it contains reliable phonetic information. Furthermore, there are models of word segmentation that rely on the stressed syllable as a segmentation cue (Cutler and Butterfield, 1992; Cutler and Norris, 1988), which implicitly highlights the importance of stress and syllables in perception.

Given the robustness of a stressed syllable, we would expect that segments in a stressed syllable are less likely to be misperceived than those in an unstressed syllable.

4.3.4 Method

All data used in the current section are the naturalistic segmental confusions, which are context-free, as described in Chapter 3, Section 3.2. Three syllable factors were computed for each of the intended segments, namely, syllable constituency (onset, nucleus, coda), syllable position (initial, medial and final), and stress (stressed and unstressed). In addition to these factors, each of the intended segments was tagged as monosyllabic or polysyllabic.

The *glmer* function from *lme4* (Bates et al., 2014) in *R* (R Core Team, 2013) was used to construct logistic mixed-effects models, with the *bobyqa* optimizer. The predictee and predictors are listed below.

Predictee: *Segment Error* (Incorrect vs. Correct)

Predictors of fixed effects: *Syllable Constituency*, *Syllable Position* and *Stress*. All the predictors are categorical.

Variables of random effects: *Intended Words*, *Utterances*, and *Corpora*.

In terms of the fixed effects, the categorical predictors need to be contrast coded. *Stress* is coded as [Unstressed vs. Stressed]. *Syllable Position* is reversed helmert coded, with [Final vs. the mean of Medial and Initial], and [Medial vs. Initial]. The reason for coding syllable position as such is to better capture the progressive effect of syllable position. Finally, *Syllable Constituency* is coded in two different ways, one for monosyllabic words and one for polysyllabic words. For the monosyllabic words, it is coded as [Onset vs. Coda] and [Nucleus vs. Coda]. For the polysyllabic words, it is coded as [Nucleus vs. Onset] and [Coda vs. Onset]. The reason for doing so will become apparent in the beginning of the analyses section.

In terms of the random effects, three variables were included, *Intended Words*, which is the intended word that contains the segment, *Utterances*, which is the

unique number given to each utterance (which is each an instance of misperception), and *Corpora*, which is the seven subcorpora used to construct the combined corpus: Browman (1978), Bird (1998), Labov (2010), Bond (Adult) (1999), Bond (Children) (1999), Nevins (2009) and Nevins (2010). These random effects would allow us to control for the variability of segment errors in specific words, utterances, and corpora.

Multiple, separate, mixed-effects models were created to test the fixed effects. They are described during the analyses.

4.3.5 Analyses

In this section, we will begin by visualising the error rates and describe any differences between error rates due to the three factors. After identifying these differences, the statistical models will then be constructed to examine whether these differences are significant. Finally, we will discuss the significant differences and whether they confirm our initial predictions of the three factors.

Before creating the mixed-effects models, we will first visualise the error rates in all possible combinations of the three syllable factors. This is done separately for monosyllabic and polysyllabic words. The visualisation is shown in Figure 4.12. The figure is divided into eight sets of bar charts. In each bar chart, there are three bars, representing the error rates of the three syllable constituents – onset, nucleus and coda – respectively. In each row of the figure, there are three bar charts, representing the error rates of the three syllable positions – initial, medial and final. The bar charts in the first two rows represent the error rates of unstressed polysyllabic words and unstressed monosyllabic words. Finally, the bar charts of the last two rows represent the error rates of the stressed polysyllabic words and stressed monosyllabic words. It is worth noting that the assignment of the syllable position for monosyllabic words as word initial is arbitrary. Interestingly, it has been suggested that monosyllables can be treated as initial syllables in terms of their phonological behaviour (Becker,

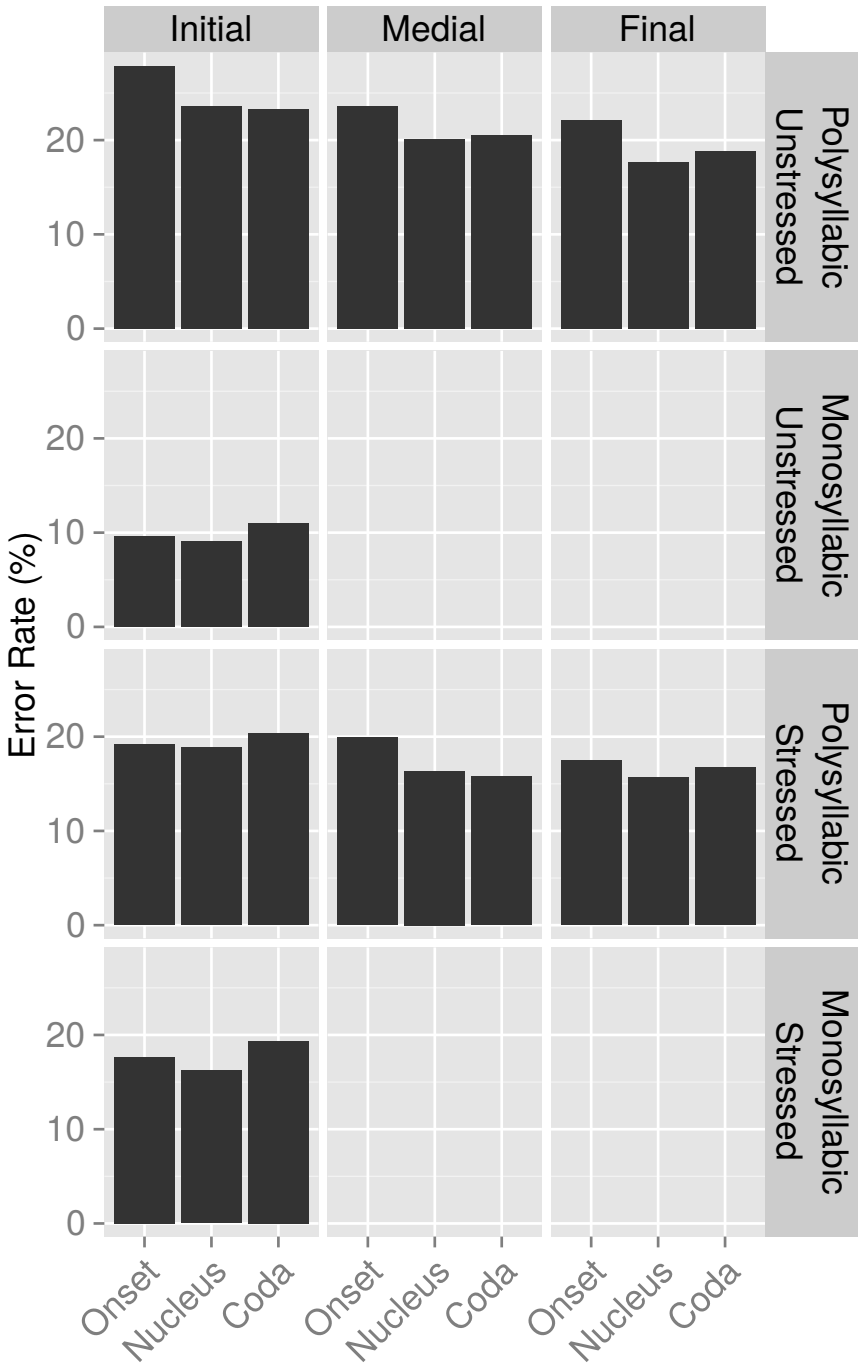


Figure 4.12: Segmental error rates by syllable constituency, syllable position, and stress: error rate is defined as the number of segmental errors in position x divided by the number of segments in position x.

Nevins, and Levine, 2012).

Starting with the unstressed polysyllabic words (the first row), we can see that there is a constituency effect and a syllable position effect. Onset has a higher error rate than nucleus and coda, regardless of the syllable position. Initial syllables have a higher error rate than medial syllables, which in turn have a higher rate than final syllables. Both of these effects can also be found with the stressed polysyllabic words (the third row), but the size of the effect seems to be weaker. The exception is that the constituency effect is absent in the initial stressed syllables in polysyllabic words. Overall, there is a stress effect in the polysyllabic words. Segments in unstressed syllables have higher error rates than those in stressed syllables.

Moving on to the monosyllabic words, there is a definite stress effect. Segments in unstressed syllables have *lower* error rates than those in stressed syllables. The direction of this effect is unexpected, and will be discussed later. Furthermore, there is a subtle constituency effect, with coda being more erroneous than onset and nucleus. This effect holds for both stressed and unstressed monosyllabic words.

4.3.5.1 Polysyllabic words

The polysyllabic words are analysed in a mixed-effects logistic model with syllable constituency, syllable position and stress as fixed effects, and intended word, utterance, and corpora as random intercepts. The model has the formula:

$$\begin{aligned} &Segment\ Error \sim\ Syllable\ Constituency + Syllable\ Position + Stress + \\ &(1|Intended\ Word) + (1|Utterances) + (1|Corpora) \end{aligned}$$

Given the syllable constituency has onset being more erroneous than nucleus and coda (Onset $>$ [Nucleus, Coda]), the following contrast coding is used to test this trend – [Nucleus vs. Onset] and [Coda vs. Onset]. If both of the contrasts are significant, then this would confirm the trend Onset $>$ [Nucleus, Coda].

Fixed effects	Estimate	SE	z	$p(>\|z\|)$
(Intercept)	-2.0202	0.1379	-14.647	$< 2 \times 10^{-16***}$
Syllable Constituency [Nucleus vs. Onset]	-0.2449	0.0341	-7.182	$6.86 \times 10^{-13***}$
Syllable Constituency [Coda vs. Onset]	-0.1288	0.0448	-2.879	0.004^{**}
Syllable Position [Medial vs. Initial]	-0.1256	0.0289	-4.353	$1.35 \times 10^{-5***}$
Syllable Position [Final vs. (Medial, Initial)]	-0.0855	0.0144	-5.956	$2.58 \times 10^{-9***}$
Stress [Unstressed vs. Stressed]	0.4152	0.0426	9.747	$< 2 \times 10^{-16***}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Random effects	Variance
Intended Word (Intercept)	2.2454
Utterances (Intercept)	1.9136
Corpora (Intercept)	0.1066

Data size	N
Observations	37,145
Intended Word	3,367
Utterances	3,621
Corpora	7

Table 4.6: Logistic mixed-effects model: predicting segment errors in stressed and unstressed polysyllabic words with syllable factors – syllable constituency, syllable position and stress.

Table 4.6 summarises the mixed-effects model. All three syllable factors – syllable constituency, syllable position and stress – are significant. Stress has the expected effect, such that a segment in an unstressed syllable is more likely to be misheard than that in a stressed syllable, as indicated by the positive estimate. Both contrasts of syllable position, [Medial vs. Initial] and [Final vs. (Medial, Initial)], have a negative estimate, which means a segment in a medial syllable is more likely to be misheard than that in an initial syllable, and a segment in a final syllable is more likely to be misheard than that in a medial or initial syllable. Both contrasts of syllable constituency, [Nucleus vs. Onset] and [Coda vs. Onset], have a negative estimate which means nucleus segments and coda segments are less likely to be misheard than onset segments.

Firstly, the mixed-effects model (as summarised in Table 4.6) confirms stress as a significant factor of segmental errors. It is in the predicted direction, Unstressed

> Stressed, with unstressed syllables being more erroneous than stressed syllables. This finding supports the idea of a stressed syllable being an "island of reliability" (Pisoni, 1981).

Secondly, syllable position is also a significant factor, and again the effect is as predicted, *Word Initial* > *Word Medial* > *Word Final*, with a decreasing amount of errors from the first syllable to the last syllable of a word. This finding can be explained by two accounts. The first account is the final lengthening effect (Lindblom, 1968; Turk and Shattuck-Hufnagel, 2007) (the amount of lengthening increases towards the end of a polysyllabic word). The second account is the predictability effect (cf. the uniqueness point Luce, 1986a) (such that the predictability of each segment increase towards the end of a word).

Thirdly, syllable constituency is also a significant factor. The effect matches one of the predictions, Onset > [Nucleus, Coda], such that onset is a more erroneous constituent than nucleus and coda. This prediction is based on the fact that there is a considerable amount of overlapping of phonetic cues between coda and nucleus, and that the rime is perceptually more salient than the onset. The fact that it does not match the prediction *Coda* > *[Onset, Nucleus]* is not too surprising; although it is based on confusion experiments, the stimuli tested were always monosyllables, and therefore the prediction should not necessarily hold for polysyllables. As we will see later, the prediction *Coda* > *[Onset, Nucleus]* is indeed more appropriate for monosyllabic words.

Finally, while both contrasts of syllable constituency are significant, the contrast [Coda vs. Onset] has a relatively high p-value of 0.004, which is perhaps due to the divergences with the initially stressed polysyllabic word condition as previously mentioned. Given the divergence, another mixed-effects model is constructed to examine whether the syllable constituency factor holds for unstressed syllables. The model has the formula:

$$Segment\ Error \sim Syllable\ Constituency + Syllable\ Position + (1|Intended\ Word) + (1|Utterances) + (1|Corpora)$$

Fixed effects	Estimate	SE	z	$p(>\|z\|)$
(Intercept)	-2.2335	0.0911	-24.518	$< 2 \times 10^{-16***}$
Syllable Constituency [Nucleus vs. Onset]	-0.0924	0.0533	-1.732	0.0834^{+}
Syllable Constituency [Coda vs. Onset]	0.0056	0.0807	0.070	$0.9443^{n.s.}$
Syllable Position [Medial vs. Initial]	-0.1351	0.0586	-2.303	0.0213^{*}
Syllable Position [Final vs. (Medial, Initial)]	0.0119	0.0358	0.331	$0.7409^{n.s.}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Random effects	Variance
Intended Word (Intercept)	2.0192
Utterances (Intercept)	2.4475
Corpora (Intercept)	0.0119

Data size	N
Observations	18,389
Intended Word	3,364
Utterances	3,619
Corpora	7

Table 4.7: Logistic mixed-effects model: predicting segment errors in stressed polysyllabic words with syllable factors – syllable constituency and syllable position.

Table 4.7 summarises the findings of the above model. Indeed both factors, syllable constituency and syllable position, are attenuated in stressed syllables. One of the two contrasts of syllable constituency [Coda vs. Onset] is insignificant; the other contrast [Nucleus vs. Onset] is only near-significant. Furthermore, one of the two contrasts of syllable position [Final vs. (Medial, Initial)] is insignificant. One explanation for such an attenuation is that there is a ceiling effect, such that the strong perceptual salience of stressed syllables overshadows the relative perceptual difference across syllable constituents and across syllable positions.

4.3.5.2 Monosyllabic words

The monosyllabic words were analysed in a mixed-effects logistic model with syllable constituency and stress as fixed effects, and intended word, utterance, and corpora

Fixed effects	Estimate	SE	z	$p(>\|z\|)$
(Intercept)	-1.6967	0.1744	-9.729	$< 2 \times 10^{-16***}$
Syllable Constituency [Nucleus vs. Coda]	-0.2649	0.0361	-7.338	$2.17 \times 10^{-13***}$
Syllable Constituency [Onset vs. Coda]	-0.1594	0.0406	-3.924	$8.71 \times 10^{-5***}$
Stress [Unstressed vs. Stressed]	-1.0064	0.0969	-10.388	$< 2 \times 10^{-16***}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Random effects	Variance
Intended Word (Intercept)	1.656
Utterances (Intercept)	2.236
Corpora (Intercept)	0.187

Data size	N
Observations	50,039
Intended Word	2,119
Utterances	4,286
Corpora	7

Table 4.8: Logistic mixed-effects model: predicting segment errors in stressed and unstressed monosyllabic words with syllable factors – syllable constituency and stress.

as random intercepts. The model has the formula:

$$\textit{Segment Error} \sim \textit{Syllable Constituency} + \textit{Syllable Position} + (1|\textit{Intended Word}) + (1|\textit{Utterances}) + (1|\textit{Corpora})$$

Table 4.8 summarises the mixed-effects model. Both syllable factors, syllable constituency and stress, are significant. Stress has an unexpected effect, such that a segment in an stressed syllable is *more* likely to be misheard than that in an unstressed syllable, as indicated by the negative estimate. Both contrasts of syllable constituency, [Nucleus vs. Coda] and [Onset vs. Coda], have a negative estimate, which means a nucleus segment and an onset segment are less likely to be misheard than a coda segment.

Firstly, the unexpected effect of stress has three potential explanations – a reporting bias in the naturalistic misperception corpora, the differing definitions of stress for monosyllables and polysyllables, and a lexical frequency effect. Recall that unstressed monosyllabic words are essentially function words in our corpus (as defined in Chapter 2, Section 2.2.3.2). Stressed monosyllabic words are therefore content

words, and they carry more information than unstressed monosyllabic words. This would mean misperceiving stressed monosyllabic words would disrupt communication more than misperceiving unstressed monosyllabic words, and therefore they are noticed and reported more frequently by the reporters of the naturalistic corpora (Browman, 1980). This is evident by the fact that in the naturalistic corpus, 543 out of 4,861 instances are misperceptions of a single word, and 535 of which are content words. Regarding the differing definitions of stress, a stressed syllable in a polysyllabic word is stressed relative to the other syllables in the same word. A stressed syllable in a monosyllabic word, however, is stressed relative to other words in the utterance (Browman, 1980). Finally, function words are of higher frequency than content words, and it is possible that high frequency words are less prone to errors; therefore, unstressed monosyllables were less erroneous than stressed monosyllables. This frequency effect is in fact confirmed in Section 4.5.

Secondly, the syllable constituency has the following effect, Coda > [Onset, Nucleus]. This is consistent with the findings from previous confusion experiments which show that coda is more erroneous than onset (Wang and Bilger, 1973; Redford and Diehl, 1999; Benkí, 2003) and that onset and nucleus are similarly less erroneous than coda (Benkí, 2003). Recall that this effect is different from the one with the polysyllabic words (Onset > [Nucleus, Coda]) but this is expected given that the previous confusion experiments were all based on the results of monosyllabic stimuli.

Interestingly, the variance of *Corpora* is the smallest of all the random effects in both mixed-effects models (Figure 4.7 and Figure 4.8). This indicates that there is a high level of consistency across corpora.

4.3.6 Conclusion

In this section, three syllable factors were examined for their effects on segmental errors in monosyllabic and polysyllabic words. The three factors were syllable con-

stituency (onset, nucleus and coda), syllable position (word initial, medial and final), and stress (stressed and unstressed).

The effect of syllable constituency in monosyllabic words is that coda segments are more likely to be misperceived than onset segments and nucleus segments – Coda > [Onset, Nucleus]. This is consistent with findings from previous confusion experiments on CV, VC and CVC nonsense syllables. This pattern can be partially explained in terms of acoustic differences between onset and coda (Redford and Diehl, 1999), in which onset is found to be acoustically more distinctive than coda. Regarding onset and nucleus having similar error rates, although Benkí's (2003) confusion data showed this pattern, no explanation was given.

Interestingly, the effect of syllable constituency is different with polysyllabic words, such that the onset is more likely to be misperceived than the rime – Onset > [Nucleus, Coda]. I argued that the mismatch is expected since the arguments and data used to support the trend with monosyllabic words were almost all based on monosyllables.

One partial explanation for this mismatch is that the true effect of syllable constituency for both monosyllabic and polysyllabic words is Coda > [Onset, Nucleus] and the mismatch is due to some of the nuclei and codas of the polysyllabic words having additional cues, which lower their error rate to a level that is even lower than the rate of their corresponding onset. Firstly, in polysyllabic words, the coda consonant in initial and medial syllables could be followed by a sonorant onset (while the coda consonant in final syllables could not), therefore getting additional transitional cues from the sonorant on the right, and these additional transitional cues can lower the error rate. Secondly, using a predictability account, the nucleus and the coda consonants in all syllable positions are predictable using their preceding segments to reduce the number of possible lexical candidates (cf. uniqueness point). This increase in predictability should be greater for the coda than the nucleus because

the coda comes after the nucleus in a syllable. Since the "true" (under this account) effect is that the coda is more erroneous than the nucleus, this greater increase of predictability for the coda would lower its high error rate.

The effect of syllable position was also confirmed, and it has the trend, Word Initial > Word Medial > Word Final. Word initial syllables are more erroneous than medial syllables, which in turn are more erroneous than final syllables. This can be explained with the final lengthening effect (Lindblom, 1968; Turk and Shattuck-Hufnagel, 2007) and/or the predictability effect (Luce, 1986a).

Both syllable constituency and syllable position effects are stronger when the syllables are unstressed rather than stressed. I argued that this attenuation in stressed syllables is due to a ceiling effect caused by the high perceptual salience of stressed syllables overshadowing both the syllable constituency and syllable position effects.

Finally, the effect of stress is that unstressed syllables are more erroneous than stressed syllables, which supports the idea that a stressed syllable is an "island of reliability" (Pisoni, 1981). The pattern, however, showed that this is only true for polysyllabic words, and not monosyllabic words. I argued that this difference is due to a reporting bias in the naturalistic corpus, differing definitions of stress between monosyllabic and polysyllabic words and/or a lexical frequency effect. Stressed monosyllabic words (basically content words) are more noticeable (and therefore reported more often) when misperceived than unstressed monosyllabic words (function words). A stressed syllable in a polysyllabic word is stressed relative to other unstressed syllables in the word, but a stressed monosyllabic word is stressed relative to the other words/syllables in the utterance. Unstressed monosyllabic words are mostly function words which are of high frequency words, and high frequency words are less likely to be misperceived.

To conclude, all three syllable factors had a definite effect on whether a segment will be misperceived. This highlights the fact that factors on the syllable level have

a top-down effect in naturalistic misperception.

4.4 Word frequency

This section examines the relationship between the frequency of the intended word and that of the perceived word. In addition, the relationship between the frequency of the intended segment and that of the perceived segment is also examined in order to eliminate the possibility that the frequency relationship between words can be reduced to the frequency relationship between segments. In other words, this is to see if the relationship of word frequency is the additive result of a lower level of frequency effect.

First, Section 4.4.1 examines the word frequency effect. Second, Section 4.4.2 examines the segmental frequency effect.

4.4.1 Word frequency

Let us start with the frequency relationship between words. Two questions are examined. First, is there a relationship between the frequency of the intended word and that of the perceived word, $Freq._{Perceived} = f(Freq._{Intended})$? Second, is the perceived word just as frequent as or more frequent than the intended word, $Freq._{Perceived} > or \approx Freq._{Intended}$ ("$>$" means more than, and "$\approx$" means similar to)?

If there is a relationship (e.g. a correlation), then we need to account for the fact that listeners can somehow estimate the frequency of the word that they were expecting, even though the actual intended word was not perceived. This estimation of the frequency of the intended word can be explained using the graceful degradation account and its extension based on the correlation between lexical frequency and duration (Vitevitch, 2002).

Graceful degradation is the ability of a processing system to not break down in a catastrophic way when the input is incomplete, but to output a representation that best matches the input (McClelland, Rumelhart, and Hinton, 1986). In the context of misperception, the perceptual system uses the information in the degraded signal to retrieve a lexical item. One of the remaining cues of the intended word could be its duration. It is based on the idea that high frequency words tend to be produced more quickly than low frequency words which tend to be produced more slowly (Wright, 1979). So although listeners cannot retrieve the intended word, the listener can still retrieve the duration of the intended word which can be used to derive the lexical frequency (Vitevitch, 2002; Tang and Nevins, 2014).

Furthermore, should we find that the perceived word has a higher frequency than the intended word, then the finding would support an account of ease of lexical retrieval. High frequency words have a lower processing cost than low frequency words and can be retrieved more quickly from the lexicon. When the intended word cannot be retrieved, listeners can either a) do their best to estimate the intended word (using its duration) and select words that can be retrieved more easily, or b) simply select words that are generally easy to retrieve, i.e. high frequency words.

4.4.1.1 $Freq._{Perceived} = f(Freq._{Intended})$

The first question is whether the frequency of the perceived word and that of the intended word have a relationship. This was addressed in previous studies using naturalistic and experimental data. Tang and Nevins (2014) conducted a similar analysis with an earlier version of the combined naturalistic corpus, which was smaller. 2,171 pairs of intended and perceived words were extracted after removing those with zero frequency and duplicates, and there was a positive correlation which is significant at a moderate level ($R = 0.33$, $df = 2{,}169$, $p < 0.002$).

In experimental studies of misperception of English words, Pollack, Rubenstein,

and Decker (1960) analysed word frequency of the intended words and the perceived words and they did not find a significant correlation. The lack of correlation could be due to the fact that the experiment tested only 144 words (which is a small sample), and the fact that the 144 words were repeatedly tested could prime the choice of the perceived words (i.e. there is a higher chance of selecting a test word as a perceived (though incorrect) word (Felty et al., 2013)). In another study, Felty et al. (2013) conducted a large word confusion experiment. 1,428 words which were randomly sampled from the English lexicon were presented in isolation with noise added to the signal. The authors found that there was a positive correlation which is significant at a moderate level ($R = 0.154$, $df = 21{,}842$, $p < 0.0001$) between the intended and perceived words. The fact that this result contradicts that of Pollack, Rubenstein, and Decker (1960) suggests that there is a subtle frequency relationship which can only be found with a larger sample.

Could the positive correlation simply be the results of confounds? Two potential confounds are identified and discussed below. The first one concerns word pairs of different lengths. Firstly, the number of syllables is usually preserved in word confusions (which constitutes 74% of the word confusion errors in Felty et al. (2013)). Therefore, long words are misperceived as long words, and short words as short words. Secondly, longer words are less frequent than shorter words. Together this means that by considering word pairs that contain words of different lengths together, a positive correlation will naturally emerge, even though there could be no (or even negative) correlation with words of the same length. For instance, the monosyllabic word pairs have a zero correlation, and the polysyllabic word pairs also have a zero correlation; but since monosyllabic words are more frequent than polysyllabic words, there will be a positive correlation when considering monosyllabic and polysyllabic word pairs together.

This would in fact explain the lack of correlation in Pollack, Rubenstein, and

Decker (1960) which only tested monosyllabic words. However, Felty et al. (2013) also tested the correlation with only monosyllabic word pairs, and a weak but significant correlation was still found ($R = 0.108$, $df = 6{,}546$, $p < 0.0001$). The fact that the correlation value dropped after controlling for the number of syllables showed that this confound is valid but nevertheless cannot fully explain all the variance. In fact, this was not controlled for in Tang and Nevins (2014), which could contribute to the significant correlation.

Another potential confound is to do with the number of identical word pairs. It is possible that a specific word is confused more often with another word; for instance, in the naturalistic data, the Labov corpus contained five instances of *copy* being perceived as *coffee*. The inclusion of these duplicate word pairs could skew the correlation if we treat the duplicates as independent data points. Tang and Nevins (2014) controlled for this by removing any duplicates, and it is not clear whether this was controlled for in Pollack, Rubenstein, and Decker (1960) and Felty et al. (2013).

In sum, the findings from both naturalistic and experimental data suggest that the frequency relationship between the intended and perceived words is stronger in naturalistic settings than in experimental settings. Furthermore, with the experimental data, the strength of the relationship appears to be dependent on experimental procedures, such as the number of stimuli and whether the stimuli were presented repeatedly. However, there are potential confounds that were not controlled for in some or all of the studies mentioned above, casting doubt on the validity of the findings.

4.4.1.2 $Freq._{Perceived} > or \approx Freq._{Intended}$

The second question concerns the nature of the frequency relationship between the intended and perceived words. This was addressed by previous studies using naturalistic and experimental data.

In naturalistic misperception, Bond (1999, p. 103) randomly sampled 75 pairs of word confusions from the author's own corpus (the Bond corpus) that have relatively simple errors and do not contain proper names. It was found that of the 75 pairs, the perceived word was more frequent than the intended word in 36 pairs, and the reverse is true in the remaining 39 pairs. This difference is not statistically significant under a chi-squared test ($\chi^2 = 0.12$, df = 1, p-value = 0.729).

In another study of naturalistic misperception, Cutler and Butterfield (1992) conducted frequency analyses of word confusions that are involved in juncture misperception, using data from the Bond corpus as well as the author's own unpublished corpus. Juncture misperception is when a word boundary is inserted or deleted, which results in one word being perceived as multiple words and multiple words being perceived as one. Starting with 246 instances of juncture misperception, 165 instances were left after removing those containing proper names or only grammatical words. The authors found that the perceived word was more frequent than the intended word in 81 pairs, and the reverse is true in the remaining 84 pairs; this difference is not significant ($\chi^2 = 0.0545$, df = 1, p-value = 0.8153).

Vitevitch (2002) re-examined this question using the Bond corpus, with a different kind of statistical analysis. The author excluded word pairs that contain complex errors. These word pairs are those with extensive mismatches as well as those which are due to juncture errors (one word is perceived as two words, and vice-versa). Furthermore, certain word pairs were excluded if the lexical variables that the author investigated (one of which is word frequency) were not available. 88 word pairs were left for analyses. An ANOVA (which is identical to an unpaired t-test) was performed using word frequency as the dependent variable and no significant differences were found. The lack of a difference would therefore imply that the intended and perceived words are similarly frequent, and that the perceived words are *not* more frequent than the intended word.

Finally, Tang and Nevins (2014) (previously mentioned) performed a similar frequency analysis. Out of the 2,171 word pairs, the number of pairs with $Freq._{Perceived} > Freq._{Intended}$ is 1,072. In the other direction, the number of pairs with $Freq._{Intended} > Freq._{Perceived}$ is 1,099. A chi-squared test yielded $\chi^2 = 0.3358$, df = 1, p-value = 0.5623, which is statistically insignificant.

Let us move on to experimental misperception. Felty et al. (2013) (previously mentioned) also analysed whether the frequency of the perceived word was significantly different from the frequency of the intended word in word confusions. It was found that the perceived words have a higher frequency than their intended words. To assess the statistical significance, instead of using the chi-squared test or t-test, a Monte Carlo simulation was done – for each word pair, the perceived word is randomly replaced with a word that has the same number of segmental differences from the intended word as the perceived word. 10,000 simulations of the word pairs were performed, and the frequency of the intended word and the fake perceived word was computed across all pairs for each simulation. They found that all 10,000 simulations have a mean difference (the perceived frequency minus the intended frequency) that is lower than the original mean difference, which suggests that the mean difference with the actual word pairs is significant.

Could these findings be explained by confounds? Three confounds are identified and discussed below. The first confound concerns duplicated pairs of word confusions, as they could skew the difference in either direction with the perceived/intended word being more frequent, and could average out any potential differences. This was controlled for in Vitevitch (2002) and Tang and Nevins (2014), but it is not clear if this was controlled for in Bond (1999, p. 103), Cutler and Butterfield (1992) and Felty et al. (2013).

The second potential confound concerns using word confusions that are involved in juncture misperception. Given that a juncture misperception involves perceiving

one word as multiple words and vice-versa, it is not clear from Cutler and Butterfield (1992) which one of the multiple words was chosen as the word for the frequency analysis. For instance, *how big is it?* was perceived as *how bigoted*. Do we take the frequency of *big*, *is* or *it*, to compare with that of *bigoted*? Given that the authors filtered out instances containing only grammatical words, it is likely that they chose the frequency of the content word, but what if there is more than one content word?

The third confound concerns using word confusions in which the intended word and perceived word are of different length (syllables or segments). Longer words tend to be less frequent than shorter words; therefore, whichever word (intended/perceived) is longer in a given pair of words, the frequency will be lower for that word. Say that on average the perceived words have fewer syllables than the intended words, then naturally the frequency of the perceived word will be higher than the frequency of the intended word. In fact, this could explain the findings by Felty et al. (2013). They found that the perceived words largely have the same number of segments and syllables as the intended words, but there is a tendency for the perceived word to be shorter (fewer segments and syllables). The fact that they found that the perceived word are more frequent can be explained by this confound. This was controlled for in Vitevitch (2002) in terms of the number of syllables, but not in Tang and Nevins (2014). It is not clear if it was controlled for in Bond (1999, p. 103) and Cutler and Butterfield (1992).

In sum, these findings from multiple studies of naturalistic misperception suggest that the frequency of the perceived word is not significantly different from that of the intended word in word confusions, i.e. they are similarly frequent. This finding is robust across the size of the sample (N = 75 – 2,171) as well as statistical methods (Chi-squared or ANOVA). Although on the surface the experimental findings contradict with naturalistic findings, the significant difference in the experimental data is perhaps confounded by the difference in word length. Again there are potential

confounds that were not controlled for in some or all of the studies mentioned above, casting doubt on the validity of the findings.

Given the potential confounds, the current analysis will re-examine the two questions while controlling for the confounds mentioned above as well as using both large and small samples of naturalistic data. To recap, the first question is whether there is a relationship between the intended and perceived word, $Freq._{Perceived} = f(Freq._{Intended})$, and the second question is whether the perceived word is more frequent than or similarly frequent to the intended word, $Freq._{Perceived} > or \approx Freq._{Intended}$. Crucially, it is possible that there is not a relationship, $Freq._{Perceived} \neq f(Freq._{Intended})$, and yet the frequency of the perceived word is still higher than the frequency of the intended word. This is the case when the perceived words are generally highly frequent, regardless of the frequency of the intended words. In the next section, I will outline the method that was used by this analysis, including the steps for evaluating the potential confounds.

4.4.1.3 Method

Using the segmental confusion data described in Chapter 3, Section 3.2, 8,259 pairs of word confusions were extracted.

Recall that the word pairs that are involved in juncture misperception can introduce complications when choosing a word pair (e.g. *big is it* > *bigoted?*). Therefore, we removed these many-to-one and one-to-many word confusion pairs, which left us with 4,268 pairs of one-to-one word confusions.

As mentioned in Section Chapter 2, 2.1.2.1, proper names are known to behave differently from non-proper names during lexical retrieval (Valentine, Brennen, and Brédart, 1996); therefore, all word confusions that involved proper names were removed. This left us with 3,244 pairs.

The token frequencies of the words found in these 3,244 pairs were extracted

from a control written English corpus as described in Chapter 2, Section 2.3. After removing the pairs containing zero frequency (i.e. not found in the corpus), 3,135 pairs remained. The token frequency was log10-transformed.

To examine the robustness of the findings, we performed our analysis repeatedly on multiple subsets of the data. The factors that are used to subset the 3,135 pairs are described below. The first factor is the choice of corpora. This is to examine if the findings are affected by certain subcorpora, since they differ in terms of collection locations, collectors' biases and sample sizes. The combined corpus, as well as the subcorpora, were considered individually (the subcorpora are described in Chapter 2, Section 2.1). The Bond corpus was divided into two, the adult misperception corpus and the children misperception corpus. The Nevins corpus was also divided into two, the data that were collected in 2009 and those collected in 2010. Therefore, together there are eight subsets (one combined corpus, and seven subcorpora), which are the Combined corpus, Browman (1978), Bird (1998), Labov (2010), Bond (Adult) (1999), Bond (Children) (1999), Nevins (2009) and Nevins (2010). The second factor is whether or not to remove duplicated word pairs. This created two subsets, those with duplicates and those without duplicates. The third factor is the removal of the word pairs with a different number of syllables. Two subsets were created, those with and without these pairs. The fourth factor concerns the number of syllables of the word pairs with matching number of syllables. Three subsets were created, the monosyllabic word pairs, the polysyllabic word pairs and those with both monosyllabic and polysyllabic word pairs.

All possible subsets of the word pairs based on these factors were tested. Should the findings be found consistently across all/most subsets then we can be doubly sure that the findings are not skewed by some or all of these factors.

Correlation analyses were performed for the question of whether the frequency of the intended word correlates with the frequency of the perceived word. A non-

parametric correlation, Spearman (two-tailed), was used to compare the two sets of frequencies, since the two sets of frequency values are not normally distributed.

Paired t-tests were performed for the question of whether the frequency of the perceived word is higher than or similar to the frequency of the intended word for a given substitution. Since the difference between two frequency values is not normally distributed, the p-values are calculated via 10,000 permutations.

4.4.1.4 Analyses

4.4.1.4.1 $Freq._{Perceived} = f(Freq._{Intended})$ Table 4.9 summarises the correlation analyses with all eight subsets of corpora, with and without duplicates. The size of the samples is shown in the columns with the header *N*. The correlation values are under the headers ρ with the level of statistical significance denoted as superscripts.

Overall, we see that all the correlation values are positive, ranging from 0.52 to 0.76, and they are highly significant ($p < 0.001$). This clearly indicates that there is a strong and significant relationship between the frequency of the intended word and that of the perceived word in word confusions. The correlation is robust across subsets of corpora, even when the sample size is small (which is the case with the Bond (Children) corpus, N = 55 or 56). It is also robust with and without duplicates, which is seen by the correlation values only dropping slightly after removing duplicates.

By visualising the correlations, we can get a better idea of the relationship and whether they are skewed by outliers. The correlations of the seven subcorpora with duplicates (excluding the combined corpus) are shown as scatterplots, each fitted with a regression line in Figure 4.13. From the figure, a strong relationship can be seen across all subcorpora, and they do not appear to be skewed/biased by extreme outliers. This indicates that the correlation values are valid.

However, the positive correlation could be due to the fact that we included both

	With Duplicates		Without Duplicates	
Corpus	N	ρ	N	ρ
Combined Corpus	3,135	0.6173^{***}	2,861	0.5767^{***}
Browman (1978)	129	0.5251^{***}	129	0.5251^{***}
Bird (1998)	259	0.7580^{***}	254	0.7466^{***}
Labov (2010)	592	0.5806^{***}	546	0.5798^{***}
Bond (Adult) (1999)	448	0.5380^{***}	440	0.5274^{***}
Bond (Children) (1999)	56	0.6011^{***}	55	0.5949^{***}
Nevins (2009)	811	0.6689^{***}	765	0.6364^{***}
Nevins (2010)	840	0.5703^{***}	815	0.5615^{***}
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.9: Correlations between the frequency of the intended word and the perceived word in word confusions, across corpora, with and without duplicates: the N columns contain the sample size and the ρ columns contain the correlation values; the superscript symbols denote the level of statistical significance.

polysyllabic word pairs and monosyllabic word pairs. Furthermore, it could be an artefact of some word pairs having a different number of syllables between the intended and the perceived words. For these reasons, we excluded the word pairs that have a different number of syllables. We then subdivided the remaining word pairs by whether they are monosyllabic or polysyllabic. We repeat this set of analyses across corpora, with and without duplicates. The correlation results, with and without duplicates, are summarised in Table 4.10 and Table 4.11 respectively. Each table shows the correlation values varied across corpora (vertically) and across subsets of syllable size (horizontally). From the left, the column *Mono. + Poly.* contains the correlations with both monosyllablic and polysyllablic word pairs, and the two on the right, *Mono.* and *Poly.*, contain the correlations with monosyllabic and polysyllabic word pairs respectively.

First of all, we examine the effect of removing pairs with a different number of syllables between the intended and the perceived words. By comparing the third column of Table 4.9 and the third column of Table 4.10, we see that the correlation values increased (though only slightly) after removing these pairs. This increase

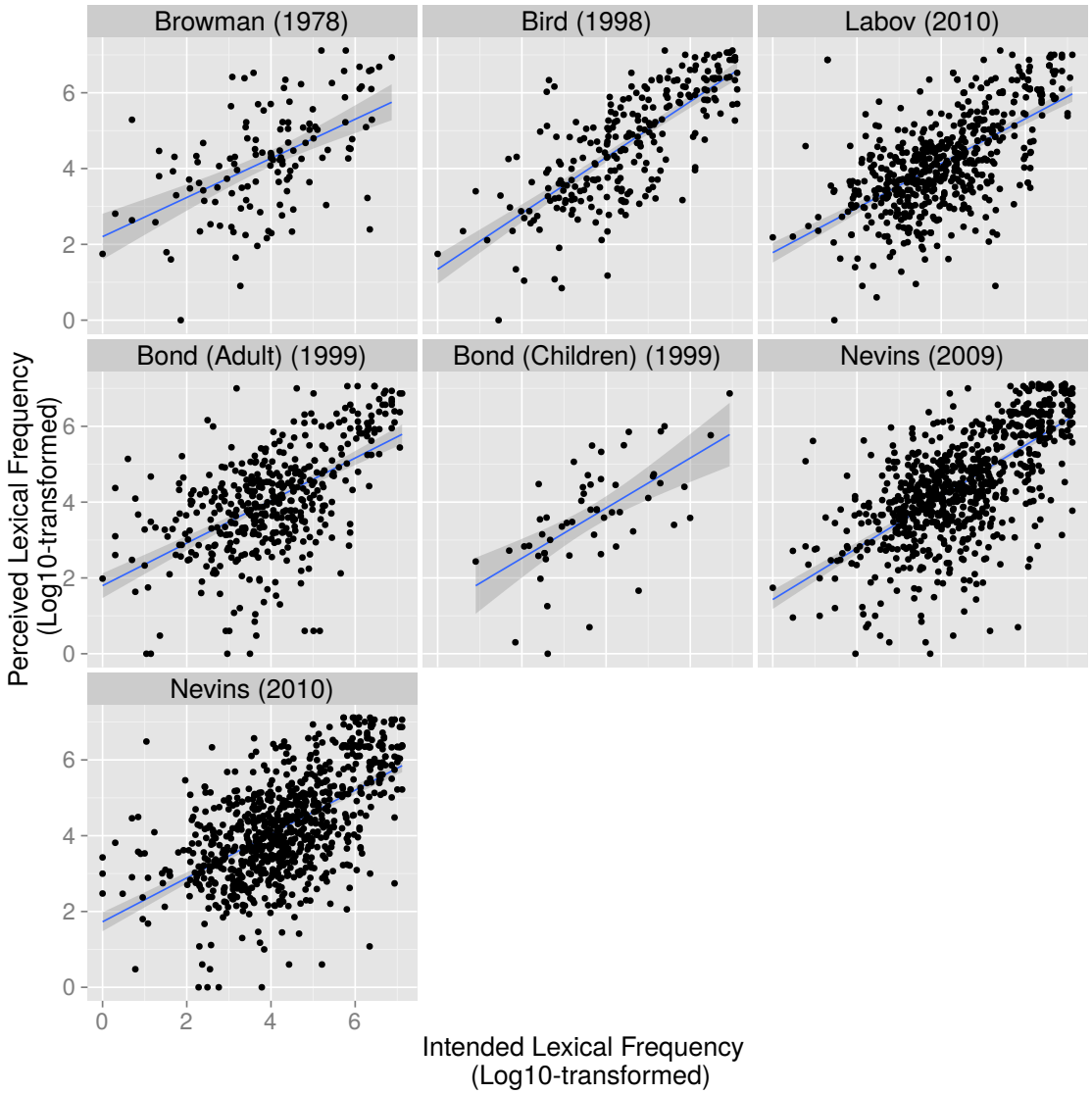


Figure 4.13: The relationship between the frequency of the intended word and the frequency of the perceived word in word confusions with duplicates, divided by corpora

is expected because these pairs have a different number of syllables, and therefore will have a larger difference in frequency. This increase is also true after removing duplicates; this can be seen by comparing the fifth column of Table 4.9 and the third column of Table 4.11.

Second, a comparison between Table 4.10 and Table 4.11, which differ in terms of whether the duplicates were removed, shows again that removing duplicates makes nearly no difference to the findings, as it only slightly lowered the correlation values.

		Mono. + Poly.		Mono.		Poly.
Corpus	N	ρ	N	ρ	N	ρ
Combined Corpus	2,668	0.6465^{***}	1,867	0.6318^{***}	801	0.3318^{***}
Browman (1978)	103	0.5593^{***}	60	0.4533^{***}	43	0.3934^{**}
Bird (1998)	223	0.7587^{***}	164	0.7082^{***}	59	0.5171^{***}
Labov (2010)	516	0.6094^{***}	366	0.6177^{***}	150	0.3094^{***}
Bond (Adult) (1999)	376	0.5962^{***}	252	0.5917^{***}	124	0.3134^{***}
Bond (Children) (1999)	51	0.5720^{***}	31	0.5471^{**}	20	$0.3729^{n.s.}$
Nevins (2009)	702	0.6973^{***}	509	0.6771^{***}	193	0.3691^{***}
Nevins (2010)	697	0.5978^{***}	485	0.5590^{***}	212	0.2646^{***}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Table 4.10: Correlations between the frequency of the intended word and the perceived word in word confusions with duplicates, subsetted by corpora and monosyllabicity: the N columns contain the sample size and the ρ columns contain the correlation values; the superscript symbols denote the level of statistical significance.

Therefore, we will not examine Table 4.11 any further.

Third, focusing on Table 4.10, we see that all but one correlation were significant. The insignificant correlation is the subset with the polysyllabic word pairs and Bond (Children) (1999) corpus which is likely due to its small sample size (N = 20). The correlations with only the monosyllabic word pairs are similarly stronger (ρ = 0.45 – 0.7) than those with both monosyllabic and polysyllabic words (ρ = 0.55 – 0.75). This agrees with Felty et al.'s (2013) findings which showed that there is a significant correlation with monosyllabic word pairs, and that the positive correlation is not merely an artefact of mixing both monosyllabic and polysyllabic word pairs. While there is a modest correlation with polysyllabic word pairs, their correlation values (ρ = 0.3 – 0.51) were nearly half as low as those with the monosyllabic word pairs (ρ = 0.55 – 0.75). This can be clearly seen in a visualisation of the correlations in Figure 4.14. The figure is divided into seven scatterplots (one for each corpus). Each scatterplot has two sets of points, one for monosyllables and the other for polysyllables, each fitted with a regression line. The difference between monosyllables and polysyllables is apparent, since the slope of the lines with the polysyllables is

	Mono. + Poly.		Mono.		Poly.	
Corpus	N	ρ	N	ρ	N	ρ
Combined Corpus	2,409	0.6020^{***}	1,634	0.5759^{***}	775	0.3180^{***}
Browman (1978)	103	0.5593^{***}	60	0.4533^{***}	43	0.3934^{**}
Bird (1998)	218	0.7460^{***}	159	0.6902^{***}	59	0.5171^{***}
Labov (2010)	477	0.6026^{***}	337	0.6166^{***}	140	0.2532^{**}
Bond (Adult) (1999)	368	0.5863^{***}	244	0.5811^{***}	124	0.3134^{***}
Bond (Children) (1999)	50	0.5650^{***}	30	0.5257^{**}	20	$0.3729^{n.s.}$
Nevins (2009)	656	0.6647^{***}	468	0.6393^{***}	188	0.3610^{***}
Nevins (2010)	673	0.5882^{***}	465	0.5440^{***}	208	0.2731^{***}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Table 4.11: Correlations between the frequency of the intended word and the perceived word in word confusions without duplicates, subsetted by corpora and monosyllabicity: the columns N contains sample size and the columns ρ contains the correlation values for each subset; the superscript symbols denote the level of statistical significance.

consistently flatter than that of the lines with the monosyllables. One explanation for this difference between monosyllabic and polysyllabic words is that the word length of the intended and perceived words was only partially controlled by matching the number of syllables, but not the number of segments. Assuming that on average each monosyllable is longer/shorter than another monosyllable by x number of segments, a polysyllabic word pair is likely to have a difference in word length of x times the number of syllables. Therefore, the difference in the number of segments between monosyllabic word pairs is likely to be smaller than that between polysyllabic word pairs. In any case, the overall relationship remains robust after dividing the pairs into monosyllables and polysyllables.

In sum, we found that there is a strong and significant correlation between the frequency of the intended word and the frequency of the perceived word in word confusions. This correlation is robust regardless of the choice of the corpus, the shape of the word (monosyllabicity), the removal of duplicates and the removal of pairs with a different number of syllables.

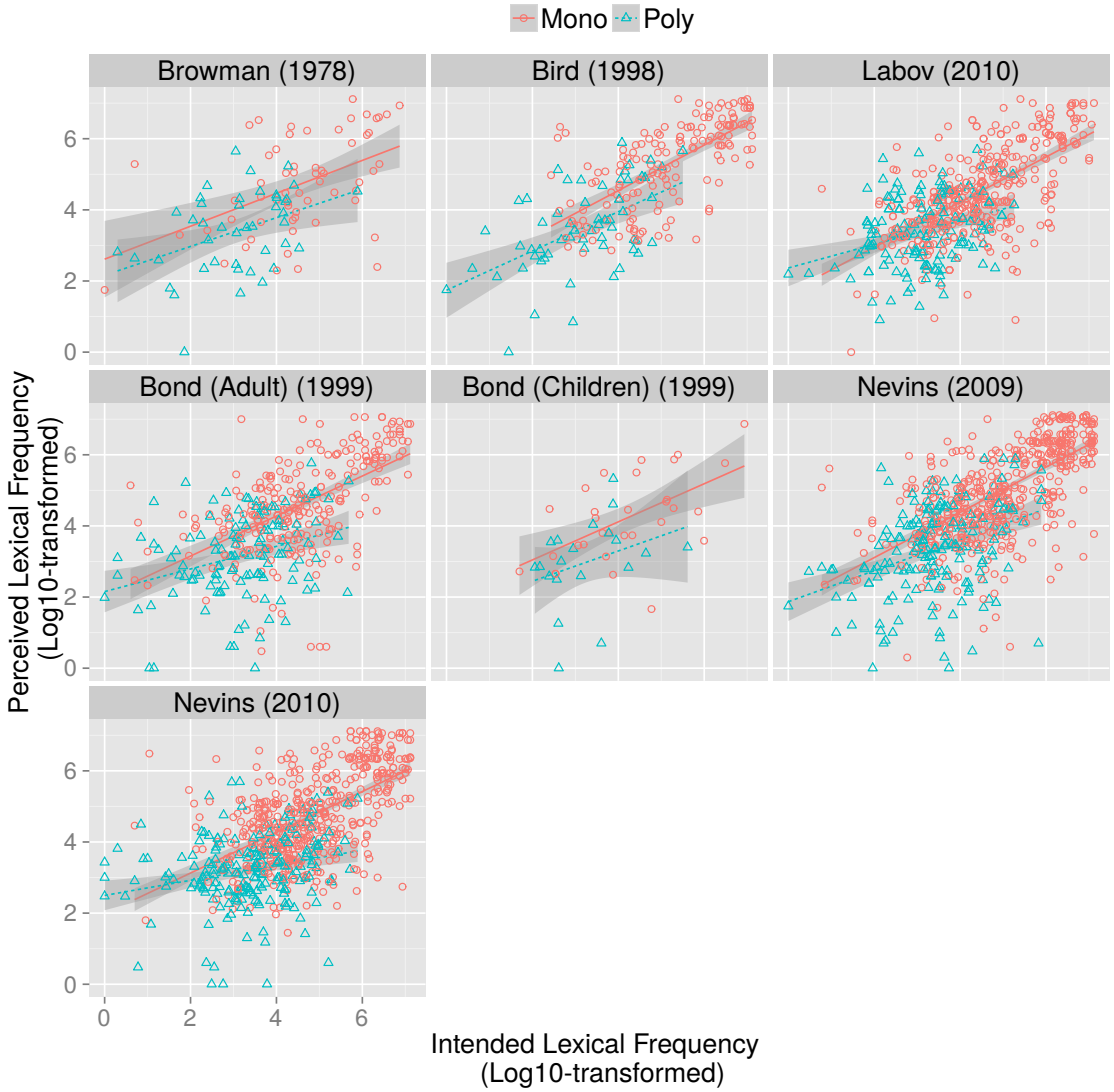


Figure 4.14: The relationship between the frequency of the intended word and the frequency of the perceived word in word confusions, with duplicates, divided by corpora and monosyllabicity

4.4.1.4.2 $Freq._{Perceived} > or \approx Freq._{Intended}$ This section examines the question of whether the frequency of the perceived word is higher than or similarly to that of the intended word in word confusions. Paired t-tests were performed for the same subsets of the word pairs as shown in the previous section.

Table 4.12 summarises the results of the paired t-tests with all eight subsets of corpora, with and without duplicates. The size of the samples is shown in the

columns with the header *N*. The t-values are under the headers *t* with the level of statistical significance denoted as superscripts as well as shown in full between the brackets. A positive t-value indicates that the frequency of the perceived word is higher than that of the intended word, and a negative t-value indicates the reverse. The bold t-values are the statistically significant ones.

Focusing on the third column (with duplicates), there is a tendency for the perceived word to be more frequent in a word confusion ($t = 0.9894$) in the combined corpus; however, it is not significant ($p = 0.1594$). Could it be that a subset of the data has a tendency in the opposite direction, thus averaging out the difference? To tackle this question, we examine each of the subcorpora. In fact, it is true that the difference is being averaged out, since the t-values of the seven subcorpora have inconsistent signs. The t-values of four subcorpora were positive, and they are Browman (1987), Bird (1998), Labov (2010) and Bond (Adult) (1999); all except the Bond corpus were significant. The t-values are negative for the remaining three, and they are Bond (Children) (1999), Nevins (2009) and Nevins (2010); only Nevins (2010) was significant. It is clear that the insignificance in regard to the combined corpus is due to the Nevins (2010) corpus and perhaps the Nevins (2009) corpus averaging out the positive t-values from the other corpora; since they are both relatively large, they therefore have a bigger effect on the combined corpus. These patterns remain the same without duplicates, as shown in the fifth column.

Furthermore, just as the correlation analyses, we analyse the effect of removing pairs that have a different number of syllables and the effect of dividing the pairs into monosyllables and polysyllables. The t-test results (subsetted by monosyllabicity and corpora), with and without duplicates, are summarised in Table 4.13 and Table 4.14 respectively.

First of all, we examine the effect of removing pairs with a different number of syllables between the intended and the perceived words. By comparing the third

Corpus	With Duplicates		Without Duplicates	
	N	t	N	t
Combined Corpus	3,135	$0.9894^{n.s.}$ (p=0.1594)	2,861	$0.6172^{n.s.}$ (p=0.2689)
Browman (1978)	129	$\mathbf{2.5299}^{**}$ (p=0.0071)	129	$\mathbf{2.5299}^{**}$ (p=0.0071)
Bird (1998)	259	$\mathbf{2.2367}^{*}$ (p=0.0124)	254	$\mathbf{2.225}^{*}$ (p=0.0139)
Labov (2010)	592	$\mathbf{2.8826}^{**}$ (p=0.0022)	546	$\mathbf{2.5977}^{**}$ (p=0.0047)
Bond (Adult) (1999)	448	$0.5081^{n.s.}$ (p=0.3023)	440	$0.4512^{n.s.}$ (p=0.3217)
Bond (Children) (1999)	56	$-0.35^{n.s.}$ (p=0.3630)	55	$-0.3628^{n.s.}$ (p=0.3594)
Nevins (2009)	811	$-0.8262^{n.s.}$ (p=0.2072)	765	$-0.8325^{n.s.}$ (p=0.2060)
Nevins (2010)	840	$\mathbf{-2.0081}^{*}$ (p=0.0215)	815	$\mathbf{-2.1356}^{*}$ (p=0.0164)
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.12: Paired t-tests (one-tailed) on the frequency of the intended and perceived words, with and without duplicates and subsetted by corpora: the N columns contain the sample size and the t columns contain the t-values; the level of statistical significance is denoted by the superscript symbols and the p-values are shown between the brackets; the bold values are the significant t-values.

column of Table 4.12 and the third column of Table 4.13, we see that the t-values decreased after removing these pairs for most of the corpora, except Bird (1998) and Nevins (2009). This decrease suggests that amongst the word pairs that have a different number of syllables there were more word pairs which have fewer syllables in the perceived word (than the intended word) than those which have more syllables in the perceived word. Nonetheless, the sign of all the t-values and the corpora, which have significant t-values, remained the same. This decrease is also true after removing duplicates, and it is big enough to change the sign of t-test with the combined corpus from positive to negative; this can be seen by comparing the fifth column of Table

Corpus	Mono. + Poly.		Mono.		Poly.	
	N	t	N	t	N	t
Combined Corpus	2668	$0.2723^{n.s.}$ (p=0.3956)	1867	$0.4297^{n.s.}$ (p=0.3319)	801	$-0.1192^{n.s.}$ (p=0.4521)
Browman (1978)	103	$\mathbf{1.7318}^{*}$ (p=0.0429)	60	$0.8733^{n.s.}$ (p=0.1966)	43	$\mathbf{1.7102}^{*}$ (p=0.0486)
Bird (1998)	223	$\mathbf{2.458}^{**}$ (p=0.0076)	164	1.5466^{+} (p=0.0585)	59	$\mathbf{2.0835}^{*}$ (p=0.0201)
Labov (2010)	516	$\mathbf{2.1364}^{*}$ (p=0.0159)	366	$0.9591^{n.s.}$ (p=0.1695)	150	$\mathbf{2.3743}^{**}$ (p=0.0092)
Bond (Adult) (1999)	376	$0.1104^{n.s.}$ (p=0.4573)	252	$0.4062^{n.s.}$ (p=0.3413)	124	$-0.3415^{n.s.}$ (p=0.3676)
Bond (Children) (1999)	51	$-0.5403^{n.s.}$ (p=0.2971)	31	$0.1427^{n.s.}$ (p=0.4446)	20	$-0.9683^{n.s.}$ (p=0.1726)
Nevins (2009)	702	$-0.5292^{n.s.}$ (p=0.2980)	509	$-0.0148^{n.s.}$ (p=0.4947)	193	$-0.8767^{n.s.}$ (p=0.1921)
Nevins (2010)	697	$\mathbf{-2.5034}^{**}$ (p=0.0071)	485	-1.4346^{+} (p=0.075)	212	$\mathbf{-2.2463}^{*}$ (p=0.0137)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Table 4.13: Paired t-tests (one-tailed) on the frequency of the intended and perceived words, with duplicates, subsetted by corpora and monosyllabicity: the N columns contain the sample size and the t columns contain the t-values; the level of statistical significance is denoted by the superscript symbols and the p-values are shown between the brackets; the bold values are the significant t-values.

4.12 and the third column of Table 4.14.

Second, a comparison between Table 4.13 and Table 4.14, which differ in terms of whether the duplicates were removed, shows again that removing duplicates makes nearly no difference to the findings. The most striking effect is that the t-value of the subset with the monosyllabic word pairs and the combined corpus is reduced from 0.4297 to near zero. Given the small difference, we will not examine Table 4.14 any further.

Third, focusing on Table 4.13, most of the t-values with the monosyllabic word pairs (the fifth column) and those with the polysyllabic word pairs (the seventh col-

Corpus	Mono. + Poly.		Mono.		Poly.	
	N	t	N	t	N	t
Combined Corpus	2,409	$-0.1227^{n.s.}$ (p=0.4532)	1,634	$7 \times 10^{-4\ n.s.}$ (p=0.4998)	775	$-0.2028^{n.s.}$ (p=0.4186)
Browman (1978)	103	$\mathbf{1.7318}^{*}$ (p=0.0411)	60	$0.8733^{n.s.}$ (p=0.1966)	43	$\mathbf{1.7102}^{*}$ (p=0.0486)
Bird (1998)	218	$\mathbf{2.4459}^{**}$ (p=0.0075)	159	1.5308^{+} (p=0.0625)	59	$\mathbf{2.0835}^{*}$ (p=0.0201)
Labov (2010)	477	$\mathbf{1.9388}^{*}$ (p=0.02605)	337	$0.827^{n.s.}$ (p=0.2096)	140	$\mathbf{2.1899}^{*}$ (p=0.0141)
Bond (Adult) (1999)	368	$0.0467^{n.s.}$ (p=0.4815)	244	$0.3259^{n.s.}$ (p=0.3762)	124	-0.3415 (p=0.3666)
Bond (Children) (1999)	50	$-0.5537^{n.s.}$ (p=0.293)	30	$0.1247^{n.s.}$ (p=0.4493)	20	$-0.9683^{n.s.}$ (p=0.1724)
Nevins (2009)	656	$-0.5343^{n.s.}$ (p=0.2964)	468	$-0.0702^{n.s.}$ (p=0.4711)	188	$-0.8047^{n.s.}$ (p=0.2144)
Nevins (2010)	673	$\mathbf{-2.6324}^{**}$ (p=0.0044)	465	-1.5905^{+} (p=0.0556)	208	$\mathbf{-2.2519}^{*}$ (p=0.0114)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Table 4.14: Paired t-tests (one-tailed) on the frequency of the intended and perceived words, without duplicates, subsetted by corpora and monosyllabicity: the N columns contain the sample size and the t columns contain the t-values; the level of statistical significance is denoted by the superscript symbols and the p-values are shown between the brackets; the bold values are the significant t-values.

umn) have similar signs to those with both the monosyllabic and polysyllabic word pairs (the third column). However, the significant t-values with both the monosyllabic and polysyllabic word pairs are significant only with the polysyllabic word pairs, and not the monosyllabic word pairs. This suggests that the overall pattern is dependent mostly on polysyllabic word pairs.

In sum, we found that the difference between the frequency of the intended word and that of the perceived word is inconsistent across corpora. The significant negative t-values with the Nevins (2010) corpus were unexpected, given that previous findings found either a null difference or a positive difference (with the perceived word being more frequent). It is not clear why Nevins (2010) and Nevins (2009) have a negative

difference, while the other corpora have a positive difference. This inconsistency suggests that the difference (positive or negative) is not robust; therefore, it is not the case that listeners would choose a higher frequency word in word confusions, but instead they would choose a similarly frequent word.

4.4.1.5 Conclusion

This section examined the two aspects of the frequency relationship between the intended and perceived words.

Firstly, we found that there is a strong and significant relationship between the frequency of the intended word and that of the perceived word – $Freq._{Perceived} = f(Freq._{Intended})$, as found in the correlation analyses. This is consistent with the large-scaled experimental confusion study by Felty et al. (2013), which also found a statistically significant correlation. Interestingly, the strength of the correlation is nearly twice as strong in our naturalistic data than the Felty et al.'s (2013) experimental data. Since Felty et al.'s (2013) study tested single words which were presented in isolation, and the naturalistic data were based on words in sentences, one might expect the frequency relationship to be stronger in Felty et al.'s (2013) study than ours. One possible explanation is that the experiment was unnatural, since in our everyday life we do not listen to words in isolation. This unnaturalness of the experiment might therefore attenuate the strength of frequency effect.

Furthermore, the polysyllabic words had weaker (but significant) correlations than the monosyllabic words, for which I have no concrete explanation. One explanation is that the difference in the number of segments between monosyllabic word pairs is likely to be smaller than that between polysyllabic word pairs, because the word length of the intended and perceived words was not matched by the number of segments.

In any case, having controlled for potential confounds, such as the choice of the

corpus, the shape of the word (monosyllabicity), the removal of duplicates and the removal of pairs with a different number of syllables, this correlation remains robust.

Secondly, we found that overall the frequency of the perceived word and the frequency of the intended word are not significantly different, which suggests that they are similar – $Freq._{Perceived} \approx Freq._{Intended}$. By subsetting the naturalistic corpus, in some of the subcorpora the perceived word was found to be significantly higher than the intended word, while in the rest of the subcorpora the direction was either reversed or insignificant. This contradicts Felty et al.'s (2013) findings that the perceived word is more frequent; however, Felty et al.'s (2013) finding could be the result of a word length confound. They found that the perceived word was on average shorter, with fewer segments and syllables, than the intended word. Given the inverse relationship between length and frequency, this naturally means that the perceived word will be more frequent. A conservative conclusion is that listeners do not simply retrieve a more frequent word.

To conclude, the findings in this section suggest that when the signal is degraded listeners would estimate the intended word with the remaining cues in the signal, such as the duration of the word. Given the relationship between frequency and duration (Wright, 1979), the resultant perceived word is therefore of a similar frequency to the intended word. Listeners do not simply retrieve an easier/more frequent word, which suggests that we should reject an ease of retrieval account.

To eliminate the possibility of this word frequency effect being the result of a segmental frequency effect, the same analyses are conducted for segments and are presented below.

4.4.2 Segmental frequency

Two questions are examined. Does the segmental frequency of the intended segment have a relationship with the segmental frequency of the perceived segment,

i.e. $Freq._{Perceived} = f(Freq._{Intended})$? Is the frequency of the perceived segment similar to or higher than that of the intended segment, i.e. $Freq._{Perceived} > or \approx Freq._{Intended}$?

Besides these two key questions, the strength of these patterns is examined between consonants and vowels, and between the three frequency measures (token frequency, type frequency and weighted type frequency) as described in Section 4.2. Three common frequency measures are examined to rule out the possibility that the strength of the relationship is strongly dependent on the chosen measure. Should we find that the segmental frequency relationship is weak (even with the best frequency measure) compared to the word frequency relationship, then it would suggest that the word frequency effect is not a by-product of the segmental frequency effect.

4.4.2.1 Method

Given that we are interested in the frequency relationship between the intended and the perceived segments in a substitution, only substitution errors are considered; and the correctly perceived segments, as well as insertion and deletion errors, were ignored. This left us with 3,329 substitution errors with the vowels, and 4,789 substitution errors with the consonants. The segmental frequency of the consonants and the vowels in the language was computed for the intended segments and perceived segments of these substitution errors.

Correlation analyses were performed for the question of whether the frequency of the intended segment is correlated with the frequency of the perceived segment. A non-parametric correlation, Spearman (two-tailed), was used to compare the two sets of frequencies, since the two sets of frequency values are not normally distributed.

Paired t-tests were performed for the question of whether the frequency of the perceived segment is higher than or similar to the frequency of the intended segment for a given substitution. Since the difference between two frequency values is not

normally distributed, the p-values were calculated via permutations. This is done with the following steps with 10,000 permutations (N = 10,000).

1. The t-value from the observed data is first calculated.
2. The data is then shuffled and a corresponding t-value is calculated.
3. The last step is repeated N times.
4. The p-value is the proportion of the absolute t-values from the shuffled data that are greater than the t-value from the observed data.

4.4.2.2 Analyses

4.4.2.2.1 $Freq._{Perceived} = f(Freq._{Intended})$ Table 4.15 summarises the correlation analyses for consonants and vowels, testing the relationship between the frequency of the intended segments and that of the perceived segments. The table shows the correlation values with their respective levels of statistical significance (as indicated by the superscripts).

	Unfiltered		Filtered	
Frequency Measure	Consonants	Vowels	Consonants	Vowels
Token	0.1631^{***}	$0.0026^{n.s.}$	0.1546^{***}	$0.0026^{n.s.}$
Type	$\mathbf{0.2032}^{***}$	0.1025^{***}	$\mathbf{0.2029}^{***}$	0.1025^{***}
Type (Weighted)	0.1868^{***}	$\mathbf{0.1109}^{***}$	0.1831^{***}	$\mathbf{0.1109}^{***}$
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.15: Segmental frequency correlations (Spearman, two-tailed) of consonants between the intended and perceived segments with three frequency measures: the superscript symbols denote the level of statistical significance; the bold value in each column is the best correlation amongst the three frequency measures.

First, we focus on the consonant correlations. In the second column, we can see that the correlation ranges from 0.16 to 0.20 across the three frequency measures, all of which are highly significant. Both measures of type frequency yield better

correlations than token frequency. Furthermore, we find the unweighted type frequency yields a better correlation than the weighted measure. Both of these findings are expected, since previous studies claimed that pattern strength in the lexicon is determined by type frequency and not token frequency (Bybee, 1995; Albright and Hayes, 2003; Hay, Pierrehumbert, and Beckman, 2004). While the correlations are significant, their strength is modest (0.1 to 0.3).

By visualising the correlations, we can have a better idea about the nature of the relationship. These three correlations are visualised as scatterplots, each fitted with a linear regression line with confidence intervals in Figure 4.15. It is immediately clear that all three plots have clear outliers at low frequency values. These outliers could inflate the correlation values. This is resolved by filtering these outliers by excluding any values that have a frequency value above or below 3 standard deviations from the mean frequency value. These filtered correlations are visualised in Figure 4.16. The gradients of the line of best fit appears to be unaffected by the filtering step. This is confirmed by their respective correlation values and levels of statistical significance as shown in the fourth column of Table 4.15. The filtered correlation values are only marginally smaller than the unfiltered ones. Together, the modest correlations (with or without extreme values) and scatterplots suggest that there is a weak relationship between the frequency of the intended segment and that of the perceived segment in substitution errors of consonants.

Moving on to the vowel substitutions, we visualise the relationship, as shown in Figure 4.17. It is clear that the slopes are flatter than those of the consonant substitutions. The slope of the token frequency is almost entirely flat, suggesting a zero correlation (i.e. no relationship). These observations are indeed confirmed in the correlations in the third column of Table 4.15. It is worth noting that the filtering step did not filter any values for vowels; therefore, the fifth column in the table is identical to the third column. Again we found that the two measures of

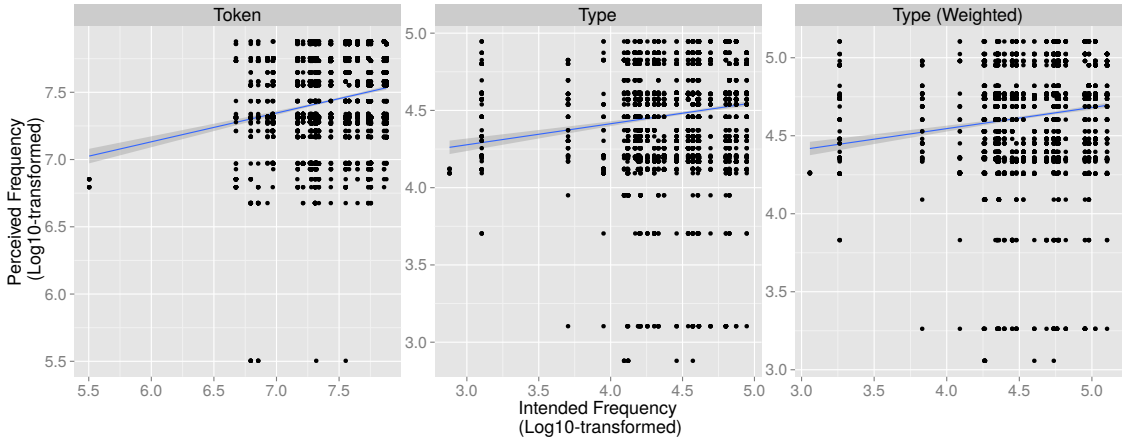


Figure 4.15: The relationship between the intended frequencies and the perceived frequencies of consonant substitutions

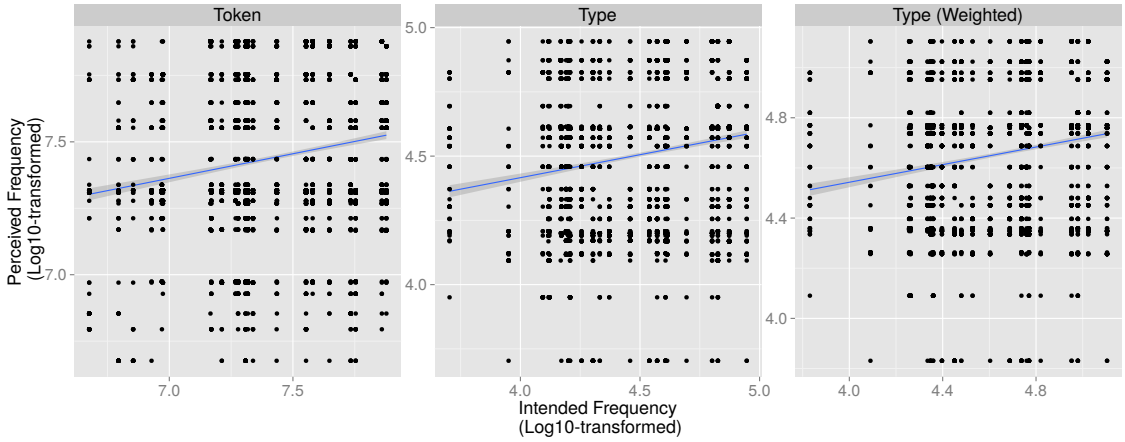


Figure 4.16: The relationship between the intended frequencies and the perceived frequencies of consonant substitutions, without extreme values

type frequency outperform token frequency in terms of the correlation values and the p-values. Crucially, the correlation with token frequency is extremely weak (ρ = 0.0026) and insignificant. Interestingly, unlike the consonant substitutions, the weighted type frequency yields a higher correlation than the unweighted measure. In any case, both the modest correlations (with the two measures type frequency) and the scatterplots suggest that there is a weak relationship between the frequency of the intended segments and that of the perceived segments for the vowels.

In sum, both consonant and vowel substitutions are governed partly by the sim-

ilarity of segmental frequency. This frequency bias is stronger for the consonants than for the vowels. The bias is sensitive to lexical information, as suggested by how the type frequency measures outperform the token frequency measure. Overall, the strength of the frequency bias is weak.

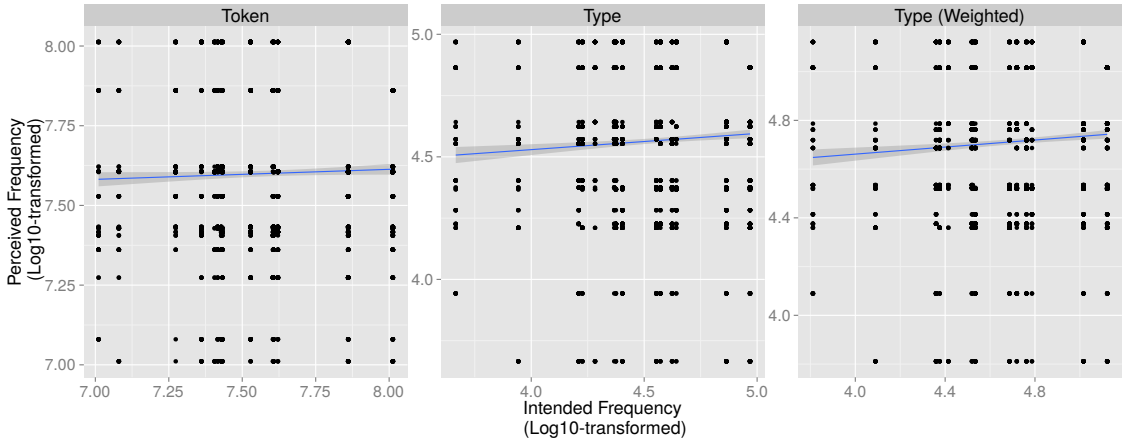


Figure 4.17: The relationship between the intended frequencies and the perceived frequencies of vowel substitutions

4.4.2.2.2 $Freq._{Perceived} > or \approx Freq._{Intended}$ The last section found that the frequency of the intended segments and that of the perceived segments are correlated. However, this does not necessarily mean that the frequency of the perceived segment tends to be higher than that of the intended segment. In this section, this question is examined.

Table 4.16 summarised the results of the paired t-tests (one-tailed), testing whether the frequency of the perceived segment is significantly different from that of the intended segment for each pair of substitutions.

The table shows that all the t-values are positive, which suggests that the frequency of the perceived segments is higher than that of the intended segments; therefore, they are not similarly frequent. However, most of the p-values were greater than 0.1, meaning that this difference is small. In terms of significance levels, only the consonants with the two type frequency measures are significant ($p < 0.05$). We

Frequency Measure	Consonants	Vowels
Token	$0.889^{n.s.}$ (p=0.1875)	$0.6061^{n.s.}$ (p=0.2714)
Type	$\mathbf{1.7687}^{*}$ (p=0.0395)	$0.6407^{n.s.}$ (p=0.2597)
Type (Weighted)	1.7625^{*} (p=0.0394)	$\mathbf{0.6563}^{n.s.}$ (p=0.2548)
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$		

Table 4.16: Paired t-tests (one-tailed) on the intended and perceived segments of consonants and vowels with three frequency measures: the level of statistical significance is denoted by the superscript symbols and the p-values are shown between the brackets; the bold value in each column is the highest t-value amongst the three frequency measures.

found that a) the two measures of type frequency yield greater t-values (i.e. a bigger difference) than the token frequency measure, b) the unweighted frequency measure yields a greater t-value than the weighted measure for the consonants and the reverse is true for the vowels, and c) the difference (across all three frequency measures) is larger for the consonants than for the vowels. These three findings were also found in the correlation analyses earlier.

In sum, there is a weak tendency for the perceived segments to be of a higher frequency than the intended segments. This tendency is statistically significant for the consonants, but not for the vowels.

4.4.2.3 Conclusion

This section examined the frequency of the intended segments and the frequency of the perceived segments that are involved in substitution errors. Concretely, we examined whether the segmental frequency of the intended segment has a relationship with the segmental frequency of the perceived segment, and whether the frequency of the perceived segment is similar to or higher than that of intended segment.

In Section 4.4.2.2.1, the analysis revealed that there is a weak relationship be-

tween the frequency of the intended segment and that of the perceived segment in substitution errors, as indicated by the modest correlation values and the corresponding scatterplots. The relationship is stronger for consonant substitutions than for vowel substitutions. Section 4.4.2.2.2 found that the perceived segment tends to be more frequent than the intended segment in a given substitution, but this tendency is only significant for the consonants.

Together, the findings suggest that the relative segmental frequencies play a minor role in consonant substitutions, and an even more minor (practically non-existent) one in vowel substitutions. The overall segmental frequency effect is weak *regardless* of the frequency measure. Comparing these findings with those in Section 4.4.1 on word frequency, it is clear that the word frequency effect cannot be reduced to a segmental frequency effect. In other words, the graceful degradation account plays a significant role in word misperception and not in segmental misperception.

4.5 Self-information

Existing models in psychoacoustic research have been developed to predict the overall speech intelligibility under the effect of noise; for instance, the Articulation Index (AI) (French and Steinberg, 1947; Steeneken and Houtgast, 1980; Rhebergen, Versfeld, and Dreschler, 2006) and Speech Intelligibility Index (SII) (ANSI, 1997). Crucially these models predict whether an utterance in a given degraded signal will be erroneously perceived.

These models can be complemented by models that predict *which* part of the utterance will be misperceived. In fact, the analyses in Section 4.3 did precisely this by using syllable factors to predict which *segment* in the word will be misperceived. Now, we will extend the size of the unit from a segment to a word by predicting which *word* in the utterance will be misperceived.

As we have examined in Section 4.4, word frequency plays a definite role in misperception in terms of lexical retrieval; the choice of the (incorrectly) perceived word is a function of the frequency of the intended word in word confusions. This section will examine a different effect of frequency on misperception, which is the predictability of word errors. In other words, if a multi-word utterance contains at least one word error, which words are most likely to be misheard?

To do so, we will examine whether the self-information of a word can predict whether a word will be misheard. By self-information, we are referring to Shannon information, which is a function of the average unpredictability in a random variable (Shannon, 1948). The Shannon information of a word, $I(word)$, is the negative log of the probability of a word, $-log(P(word))$. The probability of a word, $P(word)$, is the unconditional probability of a word, which is the number of times a word appears in a sample of the language (such as our written control corpus) divided by the total number of words in the sample. Therefore, $I(word)$ is perfectly correlated with the log of word frequency. In other words, a low frequency word has more information than a high frequency word.

In natural language, we speak in context and not with isolated words. But how do we quantify the probability of a word given its context? Since it is difficult to model the real context (e.g. world knowledge), an estimate would be to model the local context using a language model. A language model is a probability model that assigns probabilities to sentences (and indeed the words in the sentences). Using a language model, it would be possible to estimate the conditional probability of a word given the previous words, which is then converted to Shannon information, $I(word|context) = -log(p(word|context))$. That is to say, in an multi-word utterance, the self-information of a word is dependent on its preceding words. It is important to note that while the unconditional probability and the conditional probability of a word usually correlate with each other, they are not identical. In

terms of the terminology, it is worth noting that the self-information based on the conditional probability of a word given its previous words is also called *surprisal* (Hale, 2001; Levy, 2008).

What mechanisms govern the relationship between self-information and word errors? In fact, self-information plays a role in both perception and production. Each raises a different prediction on how self-information can predict word errors.

In perception, it is well-known that the processing cost of a word is a function of word frequency. That is, a high frequency word is processed faster than a low frequency word, as repeatedly demonstrated in lexical decision tasks (Brysbaert and New, 2009; New et al., 2007; Keuleers, Brysbaert, and New, 2010; Ernestus and Cutler, 2014). This is also true for the conditional probability of a word; for instance, the reading time of naturalistic texts is shorter for words that have a higher conditional probability (Smith and Levy, 2008). In other words, the processing cost is a function of the self-information of both words, $I(word)$, and words given their context, $I(word|context)$. Given this relationship between self-information and processing cost, we can expect that a word with high self-information is more likely to be misperceived because of its high processing cost.

In production, the phonetic realisation of a word is a function of its self-information. Words with high self-information tends to be spoken over a longer period (Wright, 1979; Aylett and Turk, 2004). The same applies to phonemes with high self-information, which are produced more slowly and with more articulatory detail (van Son and van Santen, 2005). More frequent (therefore less informative) words tend to undergo morphophonological reduction and alternation (Bybee, 1995; Bybee and Hopper, 2001; Bybee, 2001; Coetzee and Kawahara, 2013). Given this relationship between self-information and phonetic realisation, we would expect a word with low self-information more likely to be misperceived because of its weak phonetic cues.

Having discussed the two predictions of how self-information can predict word errors, we will describe the steps for computing the language model as well as the statistical models in the next section.

4.5.1 Method

4.5.1.1 Data selection

The combined naturalsitic corpus from Chapter 2 is used for this analysis.

All Mondegreens (misperception of music lyrics) (253 instances) and non-English misperceptions (69 instances) were excluded. The filtered corpus contains 4,861 instances of misperception. In addition, we applied two more filters. First, we removed instances that do not contain any word errors. 42 of these instances were found and removed. These instances are mostly from the Labov corpus and they consist largely of reference errors, such as the pronoun 'her' being referring to two different female entities. Second, we remove instances that contain only errors (e.g. one-word utterances). 948 instances were found and removed. The remaining instances are multi-word intended utterances with at least one word error. This left us with 3,871 instances.

4.5.1.2 Probability estimation

For each intended sentence, the conditional and unconditional probabilities of each word was estimated. The unconditional probability of a word is simply its token frequency divided by the total number of words in the control written corpus. The conditional probability of each word was computed over a language model as described below.

The probabilities were estimated by a trigram language model trained on our 353.4 million word corpus as described in Chapter 2, Section 2.3. The model was estimated using MIT Language Modeling Toolkit (v. 0.4.1) (Hsu, 2009; Hsu and

Glass, 2008), with modified Kneser-Ney smoothing (Kneser and Ney, 1995). In a trigram model, the probability of a word given its context is modelled with the probability of a word given the two previous words. If there are more than two words before a given word, then the chain rule is applied. The modified Kneser-Ney smoothing is a standard smoothing technique for trigram models (Chen and Goodman, 1999). KenLM (Heafield, 2011) was used to make queries with the model.

4.5.1.3 Statistical model

The *glmer* function from *lme4* (Bates et al., 2014) in *R* (R Core Team, 2013) was used to construct logistic mixed-effects models, with the *bobyqa* optimizer. The predictee and predictors are listed below.

Predictee: *Word Error* (Incorrect vs. Correct)

Predictors of fixed effects: $I(word)$, $I(word|context)$, *Word Length* and *Proper Name* (Proper vs. Non-Proper). All the predictors are continuous, except *Proper Name* which is categorical.

Variables of random effects: *Utterances*, *Utterance Length* and *Corpora*

In terms of the fixed effects, the two predictors of interest are the self-information of a word using its unconditional probability, $I(word) = -log(p(word))$, and that of a word using its conditional probability, $I(word|context) = -log(p(word|context))$. In addition, we included two predictors which are the controls. They are *Word Length* (estimated as the number of IPA segments in a word), and *Proper Name* (whether a word is a proper name). These control predictors were added to control for confounds, because word length is known to affect processing cost (the longer the word, the higher the cost), proper names are known to behave differently from non-proper names during lexical retrieval (Valentine, Brennen, and Brédart, 1996),

and their probability estimates might be inaccurate (e.g. the lexical frequency of *Harvard* is likely to be differ more between individuals than that of *apple*.)

In terms of the random effects, three variables were included, *Utterances*, which is the unique number given to each utterance (which is each instance of misperception), *Utterance Length*, which is the number of words in each utterance, and *Corpora*, which is the seven subcorpora used to construct the combined corpus: Browman (1978), Bird (1998), Labov (2010), Bond (Adult) (1999), Bond (Children) (1999), Nevins (2009) and Nevins (2010). These random effects would allow us to control for the variability of word errors in specific utterances, length of utterance and corpora.

To reduce collinearity, all continuous predictors, $I(word)$, $I(word|context)$ and Word Length, were standardised by scaling and centering as z-scores. The standardized predictors are henceforth referred to as $z[I(word)]$, $z[I(word|context)]$ and $z[Word\ Length]$. Following the recommendation of Rogerson (2001), any predictors with a variance inflation factor (VIF) over 5 will indicate collinearity in the model. Since all our predictors have VIF < 5 (the highest VIF was 4.97), collinearity is unlikely to be a problem.

We first fitted a model with the single terms of the predictors as fixed effects and two random intercepts.

Superset model:

$$\begin{aligned} Word\ Error \sim\ & z[I(word)] + z[I(word|context)] + z[Word\ Length] + \\ & Proper\ Name + (1|Utterances) + (1|Utterance\ Length) + (1|Corpora) \end{aligned}$$

We then performed a series of nested model comparisons on the fixed effects using ANOVA (test $= \chi^2$, $\alpha = 0.05$). The removal of terms was justified by whether a significant improvement to the model was made. If there were multiple subset models that resulted in p-values exceeding the α-level in their nested model comparisons with the superset model, the subset model with the strongest evidence (the highest p-value) was selected. We arrived at the following best model.

Best model:

$$Word\ Error \sim\ z[I(word)] + z[I(word|context)] + z[Word\ Length] + (1|Utterances) + (1|Utterance\ Length) + (1|Corpora)$$

4.5.2 Analyses

The complete summary of the best model is shown in Table 4.17. The model suggests the following significant predictors: $z[I(word)]$, $z[I(word|context)]$ and $z[Word\ Length]$. The predictor *Proper Name* was dropped during the nested model comparison, indicating that it makes an insignificant contribution to the model, and as such it is not a useful predictor of word errors.

In the table, a positive estimate means that the corresponding predictor has a positive relationship to the likelihood of a word error, and therefore a negative estimate means that there is a negative relationship. Given that we have converted our continuous predictors to z-scores, the absolute values of the estimates are now comparable between predictors.

From the table, the strongest to the weakest predictors are $z[I(word)]$, with an estimate of 1.1247, the control predictor $z[Word\ Length]$, with an estimate of -0.4144, and $z[I(word|context)]$, with an estimate of 0.2764.

Firstly, both measures of self-information survived the nested model comparison. This means that they are both important predictors of whether a word will be misheard. Secondly, the signs of the estimates of the $z[I(word)]$ and $z[Word\ Length]$ are both positive. This supports the prediction that there is a relationship between processing cost and self-information, such that a word with high self-information is more likely to be misheard. The prediction that word errors are dependent on phonetic reduction (due to low self-information) is rejected.

Secondly, the control predictor, $z[Word\ Length]$, has an estimate of -0.4144, which suggests that the longer the word is, the less likely the word would be misper-

ceived. This is consistent the findings in previous studies such as Wiener and Miller (1946) and Felty et al. (2013). Given that the word confusions in Felty et al. (2013) were induced by presenting participants with words in isolation, the fact that the word length effect is also found in our naturalistic corpus suggests that the length effect is robust not only in isolation but also in words that are presented together with other words.

Finally, R^2_{GLMM}, the percentage of variance explained by the model, is calculated (Nakagawa and Schielzeth, 2013; Johnson, 2014; Bartoń, 2014) with marginal R^2_{GLMM} being 24% and conditional R^2_{GLMM} being 34%. Marginal R^2_{GLMM} represents the variance explained by fixed effects and conditional R^2_{GLMM} represents the variance explained by both fixed and random effects. The difference between conditional R^2_{GLMM} and Marginal R^2_{GLMM} indicates that the random effects did capture a sizeable portion (10%) of the variance. From Table 4.17, we see the variance of the *Utterance Length* is the highest of all three random intercepts, and therefore it contributes most towards the 10% of the variance. In other words, there is a considerable amount of variation in predicting word errors that depends on the length of the utterance. Interestingly, the variance of *Corpora* was 0.0718 which is a lot lower than that of *Utterance Length*. This indicates that there is a high level of consistency across corpora. Most importantly, the fixed effects capture twice as much variance as the random effects, highlighting the strong relationship between the fixed effects and the likelihood of word errors.

4.5.3 Conclusion

This section examined the effect of self-information on word errors. Two types of self-information were tested. The first type is based on the unconditional probability of a word, namely token frequency – $I(word)$. The second type is based on the conditional probability of a word given its preceding context, $I(word|context)$.

Fixed effects	Estimate	SE	z	$p(>\|z\|)$
(Intercept)	-1.7563	0.1867	-9.410	$< 2 \times 10^{-16***}$
$z[I(word)]$	1.1247	0.0394	28.540	$< 2 \times 10^{-16***}$
$z[I(word\|context)]$	0.2764	0.0315	8.786	$< 2 \times 10^{-16***}$
$z[Word\ Length]$	-0.4144	0.0269	-15.410	$< 2 \times 10^{-16***}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$

Random effects	Variance
Utterances (Intercept)	0.0475
Utterance Length (Intercept)	0.3775
Corpora (Intercept)	0.0718

Data size	N
Observations	19,840
Utterances	3,739
Utterance Length	26
Corpora	7

Table 4.17: Best logistic mixed-effects model: predicting word errors with self-information

Having controlled for potential confounds, such as word length and whether the word is a proper name, and allowed for variations in corpora and utterances, both types of self-information were still strong and significant predictors of whether a word will be misperceived in an utterance. The amount of self-information of a word is positively related to the likelihood of a word error. In other words, the less predicable a word is, the more likely it is that it will be perceived. This confirmed the processing cost account, in the respect that high self-information words have a higher processing cost. The findings also rejected the phonetic reduction account, due to the words being phonetically more reduced if they have low self-information, and as such phonetically reduced words are hardly to perceive.

4.6 Conclusion

The focus of this chapter was to examine the top-down lexical factors that play a role in naturalistic misperception. Top-down factors were tested from linguistic units of

various sizes – segments, syllables, words, and utterances – and the strength of their effects was evaluated.

Section 4.2 examined the effect of segmental frequency on two different aspects of segmental confusions. Firstly, the target bias and the response bias are strongly dependent on segmental frequency as confirmed by the strong to very strong correlations. This frequency bias is true for substitutions, insertions and deletions. In a segmental misperception, the probability of a given segment being the target (the intended segment) or the response (the perceived segment) is dependent on the probability of this segment occurring in the language, i.e. its frequency. The more frequent a segment is, the more likely it is that it will be the intended segment that gets misheard, and the perceived segment that is the result of a misperception.

Secondly, the difference in segmental frequency was found to be a significant predictor of the direction and strength of the asymmetrical confusions. For a given pair of segments, the confusion pattern tends to be in the direction of the more frequent segment, such that the less frequent segment is perceived as the more frequent segment more often than the reverse.

In Section 4.3, the three syllable factors – syllable constituency (onset, nucleus, and coda), syllable position (initial, medial, and final), and stress (unstressed and stressed) – were found to have a definite effect on the likelihood of a segment error. However, the effect of syllable constituency and that of stress are different between monosyllabic words and polysyllabic words. In monosyllabic words, coda is more erroneous than both onset and nucleus – Coda > [Onset, Nucleus]. This pattern is consistent with the findings from previous confusion experiments (Wang and Bilger, 1973; Redford and Diehl, 1999; Benkí, 2003).

In polysyllabic words, the constituency pattern is, however, different, with onset being more erroneous than both nucleus and coda – Onset > [Nucleus, Coda]. This is robust across syllable positions and stress conditions. I argued that this mis-

match between monosyllabic words and polysyllabic words is expected, since all the external evidence was based on monosyllables. One explanation was proposed using arguments of predictability and additional transitional cues. Segments become more predictable incrementally towards the end of a word, and this effect should be stronger for polysyllabic words than monosyllabic words, given the findings on uniqueness point (Luce, 1986a). Codas in word initial and medial syllables could have extra transitional cues from sonorant onsets on the right. Assuming that the true effect is the one found with monosyllabic words, Coda > [Onset, Nucleus], both of these factors could decrease the error rates of coda and nucleus and as a result onset becomes more erroneous.

For polysyllabic words, stressed syllables were less erroneous than unstressed syllables, which supports the idea that stressed syllables are "islands of reliability" (Pisoni, 1981). However, the effect was reversed for monosyllabic words, which can be explained in three different ways – a reporting bias, a differing definition of stress between monosyllabic and polysyllabic words, and lexical frequency.

Syllable position has a robust effect, in that word initial syllables are more erroneous than medial syllables, which in turn are more erroneous than final syllables. This effect is stronger in unstressed syllables and is attenuated in stressed syllables. This attenuation was argued to be a result of a ceiling effect caused by the high perceptual salience of stress overshadowing other factors.

Section 4.4 examined the frequency relationship between intended and perceived words. A strong and significant relationship was found between the frequency of the intended word and that of the perceived word – $Freq._{Perceived} = f(Freq._{Intended})$. This relationship remained robust after controlling for potential confounds, such as the choice of the corpus, the shape of the word (monosyllabicity), the removal of duplicates and the removal of pairs with a different number of syllables. This is consistent with the large-scale experimental confusion study by Felty et al. (2013).

The frequency of the perceived word and the frequency of the intended word are not significantly different in the combined corpus, and the pattern is inconsistent across subcorpora. This suggests that there is not a robust frequency difference – $Freq._{Perceived} \approx Freq._{Intended}$. In addition, the frequency relationship between the intended and perceived words cannot be reduced to that of the intended and perceived segments, as indicated by the weak correlation between the frequency of the intended segment and that of the perceived segment in substitution and the inconsistency between consonants and vowels — the frequency of the perceived segment is significantly higher than that of the intended segment only for consonants, and not for vowels.

Section 4.5 evaluated the effect of self-information on the likelihood of a word error in an utterance. After controlling for word length, the conditional and unconditional self-information of a word were both strong and significant predictors of word errors. High self-information words were more erroneous than low self-information words. This supports the processing cost account, which states that words that are harder to retrieve/process are more erroneous. The findings also suggest that listeners are sensitive to the conditional probability of a word, as opposed to only the unconditional probability (i.e. token frequency) when processing speech on an utterance level. This highlights the fact that in naturalistic speech perception, we do not process words in isolation, but in context. This casts doubt on the ecological validity of confusion studies which present words in isolation (Cooke, 2009; Felty et al., 2013; Tóth et al., 2015).

To conclude, the four sets of analyses in this chapter have demonstrated that naturalistic misperception is dependent on top-down factors from a range of linguistic units – segments, syllables, words, and utterances. This complements our findings in Chapter 3 in which phonological and phonetic factors were found. On the whole, we successfully replicated findings by Bird (1998), Browman (1978), Bond (1999),

Vitevitch (2002), and Tang and Nevins (2014) which used much smaller sets of naturalistic data. Crucially, this chapter presented, for the first time, a wide range of analyses of top-down factors using the largest naturalistic corpus of misperception.