Measuring segmental and lexical trends in a corpus of naturalistic speech misperception*

Kevin Tang & Andrew Nevins

University College London

1. Introduction

Slips of the ear are generally agreed to be speech misperceptions of an intended speech signal (Bond 1999).[1] The word "intended" is important here, as slips of the ear are not speech misproductions where the mismatch lies between the intended utterance and the actual utterance: the mismatch lies between the produced utterance (by a speaker) and the perceived utterance (by a listener). For example, a speaker might intend to produce "doll", and successfully produce "doll", but a hearer perceives "doll" as "dog". Such errors can be revealing of a number of natural tendencies, and arguably furnish the short-term 'mutations' that, when recurrent enough, fuel language change on longer timescales. As Laver (1970, p.61) put it, "The strategy of inferring properties of control systems from their output can be applied not only to the efficient operation of these systems, but also to their output when malfunctions occur. The evidence from characteristic malfunctions is more penetrating than that obtained when the system is operating efficiently."

While a large amount of laboratory-based misperception studies exist (e.g. Miller and Nicely (1955) et seq.), these use isolated nonsense syllables, devoid of conversational context. Recent work has started to explore misperceptions experimentally on a word-level, e.g. Cooke et al. (in press) and Lecumberri et al. (2013). An open question, therefore, is whether the same perceptual trends hold within naturalistically occurring misperceptions, especially given the influence of lexical factors and top-down conversational influences. The amount of research and data on naturally occurring slips of the ear is scarce. The first known work on slips of the ear was Meringer (1908), which was based on a corpus of 47 slips in German. This work subsequently inspired other collections, namely Browman (1980) (≈200 slips), Labov (2010, chap. 2) (≈900 slips), and perhaps most noticeably, Bond (1999), containing approximately 900 slips. Even though Bond (1999)'s study lacked

*We would like to thank the audience at NELS43, A. Cutler, J. Harris, M. Becker and K. Abels.

[1]For an overview of the interest of slips of the ear and mondegreens to cognitive science, see [https://www.youtube.com/watch?v=a_ejEzqkkdQ] and [https://www.youtube.com/watch?v=dBnhkwRmYuQ]

detailed quantitative analysis, it did demonstrate that these slips are important linguistic phenomena, and can be classified among a number of dimensions, thereby calling for the need to perform a quantitative analysis.

1.1 Complementarity of experimental data and diary corpora

The advantage of naturalistic misperceptions is their authenticity, though it is nearly impossible to obtain speech recordings of their occurrence. Lecumberri et al. (2013, p.1) remark that "Misperceptions occurring in naturalistic settings are clearly most authentic, but the speech − and, equally-importantly, the misperception-inducing context − is almost never recorded at the signal level in a form suitable for further analysis and replication with other listeners." Although the data of diary corpora have potential reliability issues, as argued extensively by Cutler (1982) and Ferber (1991), the counterarguments and successes in using diary corpora to support experimental data have been overwhelming (Cutler and Butterfield 1992, Bond 1999, Vitevitch 2002, Labov 2010, Hirjee and Brown 2010). We therefore argue that diary data and experimental data are complementary, and that naturalistic data remain highly valuable. The arguments for and against diary corpora are discussed below.

Arguments against diary corpora usually center around the fact that they are observational data from uncontrolled sampling. Cutler (1982) provided a detailed discussion on the reliability of data of this kind with a focus on misproduction, concentrating on the issue of detectability of different slip types which is dependent on factors such as hearing errors, perceptual confusions and others. A potential confound for speech misperception is that they could in fact be misproduction. Another relevant confound is the issue of reconstruction. Consider that when the hearer perceives an implausible word, the hearer could reconstruct a plausible word (possibly even the same word as the intended word) as a repair strategy. These misperception instances would not be recorded by the reporters, and there is no way of knowing how often this occurs, thus potentially biasing the data.

Ferber (1991) further argued against naturalistic data particularly by highlighting the reliability of the collection process itself. The study provided a list of possible factors that could affect the quality of the data. The reasons were that naturalistic slips might not be recognized as such, might not be remembered, might occur too frequently to record, be incorrectly transcribed, or be recorded with insufficient context. Ferber's study examined the consistency of on-line slip collection with three people listening to the same recording. The results suggested that the consistency between collectors was low, and that there was not a trend of particular types of slips being more detectable than others.

On the other hand, experimental findings often show a high rate of agreement with naturalistic findings, for example in Cutler and Butterfield (1992), Bond (1999), Vitevitch (2002) and Labov (2010, chap. 3 and 4), as discussed below.

Cutler and Butterfield (1992) set out to test the rhythmic segmentation hypothesis, which predicts that English listeners would operate on the assumption that strong syllables are likely to be the initial syllable, while weak syllables are either not word-initial and if they were, they will likely to be grammatical words. The study started with an analysis of 246 juncture misperceptions from Bond (1999)'s corpus, and found that there are indeed more juncture insertions before strong syllables than weak syllables, and more juncture

deletions before weak syllables than strong syllables. By using faint speech in an experiment, where the speech input is presented faintly at the level at which listeners could hear about 50% of the input, the same error pattern was found in listeners' misperception.

In Music Information Research (MIR), the 'misheard lyric matching problem' has been known to be an challenge for internet search engines. This is when users enter misheard lyrics in a search engine, hoping to find the intended lyrics. One model using diary corpora was introduced by Hirjee and Brown (2010). They utilized diary corpora of speech misperception of lyrics rather than of conversational speech. The data were 20,788 instances from misheard lyrics websites (e.g. `kissthisguy.com`) that collect these instances from the public. They introduced a probabilistic model of mishearing by calculating phoneme confusion frequencies of the misheard lyrics. Their model, when tested on 146 misheard lyrics queries, was able to find up to 8% more correct lyrics than other methods that did not use these naturalistic data, such as phoneme edit distance. This study suggests that by using diary corpora of misperception, they were able to better predict people's perceptual errors of lyrics, thus strengthening the importance of naturalistic data.

2. Methodology

In order to quantitatively measure whether certain segmental and lexical factors recur to a significant extent within a large corpus of slips of the ear, a number of steps are required. The overall methods are to first phonetically transcribe the intended and perceived utterances, and then to feed the pairs of intended and perceived sequences into an alignment algorithm. Finally, the output are confusion matrices of substitutions, deletion and insertions of segments, with which lexical errors can also be identified.

2.1 Data collection

Naturalistic, diary data has a number of advantages. The apparent disadvantages can often be overcome by collecting data in sufficiently high numbers and in ensuring consistent transcription. Previous efforts exist for error data along these lines. Focusing on slips of the tongue, Dell and Reich (1981) recruited about 200 linguistics students in five one-month periods between 1975 and 1977 and collected 4000 instances of speech misproduction.

Following Dell and Reich (1981)'s footsteps, in an effort to collect a large number of slips of the ear through the efforts of a wide range of trained students, one of us (Nevins) recruited 24 linguistics students at Harvard University in a course about speech misperception, where students were made aware of the various kinds of errors for one semester per year in 2009 and 2010. They were instructed to record the intended and the perceived utterances, their phonetic transcriptions, the demographics of the speakers and hearers (age, gender, accent, native and non-native language(s) and hometown) and finally the context of each misperception, and any comments or corrections by interlocutors, including, where possible, a summary for how the slip was detected.

This collection, at the end of the two years, yielded 2857 misperception instances in American English, of which 1523 instances were collected in 2009, and 1334 instances in

2010. At the moment, these two corpora are larger than any of the existing corpora (see Figure 1 for a snapshot of Nevins' 2010 corpus).

Figure 1: A snapshot of Nevins' 2010 corpus

Intended	*Perceived*	*IPA Intended*	*IPA Perceived*	*Utterer*	*Perceiver*	*Where*	*Collected By*	*Notes*	*Topic of Conversation*	*Request for clarification*
Fill it	Willis	| | ˈfɪl ɪt | |	| | ˈwɪl.ɪs | |	Male, 19, MD	Female, 18, NY	Annenberg	Kristen	Lots of background noise	Waterbottles	You named it?
Think about it	Think about shit	| | ˈθɪŋk ə.ˈbaʊt ɪt | |	| | ˈθɪŋk ə.ˈbaʊt ˈʃɪt | |	Female, 18, NY	Male, 18, TX	Annenberg	Kristen	Lots of background noise	Life Sci	Haha What?
Fire her	Fire	| | ˈfɑɪ.əɹ əɹ | |	| | ˈfɑɪ.əɹ | |	Female, 18, NY	Female, 18, CT	Annenberg	Kristen	Lots of background noise	Politics	Huh?
Smart	Start	| | ˈsmɑːɹt | |	| | ˈstɑːɹt | |	Female, 18, NY	Male, 19, MD	Lamont	Kristen	Whispering	Homework	I already did.
It was hot as balls.	I was sucking balls.	| | ɪt wəz ˈhɔːt æz ˈbɑːlz | |	| | aɪ wəz ˈsʌk.ɪŋ ˈbɑːlz | |	Male, 18, LA	Male, 18, CO	Annenberg	Preston	Background noise	A heated situation	You what?
I need to have fun tonight because I got a job.	I need to have fun tonight because I got into Harvard.	| | aɪ ˈniːd tə ˈhæv ˈfʌn tə.ˈnaɪt bɪ.ˈkɔːz aɪ ˈgɑːt ə ˈdʒɑːb | |	| | aɪ ˈniːd tə ˈhæv ˈfʌn tə.ˈnaɪt bɪ.ˈkɔːz aɪ ˈgɑːt ˈɪn.tu ˈhɑː.ɹv.əɹd | |	Male, 18, CO	Male, 18, Alberta, Canada	Annenberg	Preston	Background noise	The gain of a job	What?
Is Turkish Balto-monosyllabic?	Is Turkish Balto-Slavic?	| | ɪz ˈtɜːɹk.ɪʃ ˌmɑːn.ə.sɪ.ˈlæb.ɪk | |	| | ɪz ˈtɜːɹk.ɪʃ ˌbɑːlt.ə.ˈslæv.ɪk | |	Male, 18, IL	Male, 18, CO	Memorial Church	Preston	Reverberation	Turkish's linguistic context	none
I want to take	I want to take a	| | aɪ ˈwɑːnt tə ˈteɪk ə ˌviː.iː.ˈɛs	| | aɪ ˈwɑːnt tə ˈteɪk ə ˌbiː.ˈɛs	Female,	Male, 18,	Baskin		Background	Courses for	

2.2 Compilation of five different corpora into one

To increase the overall size of the data, three existing corpora, Browman (1980), Bond (1999) and Labov (2010, chap. 2), of a reasonable size were combined with the two large corpora by Nevins. Like many error corpora, the data in Browman (1980) and Bond (1999) were not published digitally, though fortunately they are published in the appendices of the publications. The appendices were scanned and put through optical character recognition. The $\approx$ 900 instances were then manually checked against the hard copy. Regarding (Labov 2010, chap. 2), the author kindly provided us with the full corpus digitally. The total number of misperception instances amounts to $\approx$ 5000, which is the largest corpus to our knowledge. Much work was put in to compiling all five corpora and standardizing their format. Many of the reported data required substantial cleaning, e.g. spell-checking, separating the demographic data into usable parts, standardizing the phonetic transcriptions (see Section 2.3 for details) and more. It is currently being cross-verified, permissions are being sought, and it is being converted into easily searchable formats, with the aim to make it publicly available within the future.

2.3 Transcription

Transcription of this data was a challenging and time-consuming part of the analysis. The first issue in the transcription stage arose out of the fact that a high number of misperceptions were pairs of sentences rather than pairs of words, and the reporters tended to transcribe only the words that were affected, and left the rest of the sentence untranscribed. The question is whether a word-level, as opposed to sentence-level, transcription is sufficient or not, especially given the possibility of sandhi phenomena. A sentence-level transcription would be extremely useful for context-sensitive analysis and for estimating the relative *rate* of errors by normalizing with the number of intended segments.

The second issue was that the misperceptions were reported by approximately two dozen unique reporters whose transcription styles varied along certain specific dimensions, e.g. the use of different phonetic notations; the choice of narrow/broad transcriptions, and the choice of vowels, e.g. [ə]-[ʌ], [ɛ]-[e], [eɪ]-[ej] etc.. Fortunately, due to the detailed nature of the corpus, the speakers'/listeners' demographics were available for deciphering the reported vowels, and reporters' names were used to decipher any inconsistencies.

The Longman Pronunciation Dictionary (LPD) (Wells 2008) and the Longman Dictionary of Contemporary English (Fox and Combley 2009) were used for checking the reported transcriptions and stress patterns of compounds. IPA was chosen to be the convention in the transcription of this corpus along with primary and secondary stresses, syllable breaks and word boundaries. A set of basic transcription conventions was devised for the current purpose: unless indicated by the reported transcriptions, 1) all content words and polysyllabic function words were stress-marked and the others were not; 2) lexical and rule-governed stresses are considered, and in cases where the stress shift rules are optional, the intended rules were only applied if the result would create a match between the stress patterns of the intended and perceived utterances; 3) the weak forms of function words were preferred; 4) since LPD contains variations of pronunciation and preference polls are available for some entries, the choice of the variations was based on the demographics of the speakers and listeners.

2.4 Alignment

The key aspect of analysing the misperception data is to identify the change. However, there are at least two potential difficulties, namely the amount of data, and the identification of changes. Slips containing only one 'error' - an insertion, deletion or substitution - can be analyzed rather straightforwardly, e.g. in 'thug' → 'hug', [θʌg] → [hʌg], the change is [θ] → [h]. However it is not so simple with multiple error slips, e.g. 'sleeping bag' → 'single man', [sliːpɪŋ bæg] → [sɪŋgəl mæn], which have many possible analyses or alignments, and therefore it is clear that the complexity of the analysis increases with the number of errors per slip. Manual alignment by visually identifying changes is infeasible for large corpora such as ours. Furthermore, manually aligning slips would be a subjective process and the quality would vary depending on the analyst and his/her judgement. The need for a computational analysis is apparent, since it is both automatic and objective.

In molecular biology, many methods have been developed to align DNA sequences, and these methods or algorithms can be adapted for phonetic alignment purposes. A mainstream algorithm, the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), was used.

The Needleman-Wunsch algorithm involves a gap penalty which can be constant, linear and affine. Different gap penalty schemes require different modifications to the algorithm. The affine gap penalty scheme uses two parameters, the cost of a gap opening and the cost for a gap extension, which is a function of the gap length l, and thus is either $GapPenalty = g_{open} + lg_{extend}$ or $GapPenalty = g_{open} + (l-1)g_{extend}$ (Eidhammer et al. 2004). This scheme can favor big gaps over many smaller gaps of the equivalent size, or vice-versa. It was chosen for this study because it could be beneficial for capturing slips that involve whole-word deletions, rather than just simple isolated-segment deletions.

In linguistics, phonetically-based alignments have been developed for aligning pairs of cognates, bilingual texts, speech misproduction data and more. One example is called ALINE (Kondrak 2003), which uses phonetic similarity to improve alignment accuracy by defining multiple distinctive features such as voice, lateral, place, nasal and others and relies on the assumption that similar sounds are more likely to correspond to each other. In fact, Browman (1980)'s alignment algorithm was also a phonetically-based one using phonological features.

On the one hand, it is unclear whether using a phonetically-based alignment algorithm would be beneficial for this analysis. Considering that the aim is to find segmental changes and how phonetic similarity motivates these changes, a phonetically-blind alignment algorithm might be more suitable for the current analysis rather than a phonetically-based one in order to avoid any potential circularity. On the other hand, if the algorithm is phonetically-blind, a consonant-consonant substitution would be penalized as much as a consonant-vowel one, and this is also highly undesirable.

Motivated by how Browman (1980) manually aligned the syllables between the intended and perceived utterances and the assumption that syllables are most likely to be preserved in misperception and the conservation of syllable count, a good compromise was to use an algorithm which is phonetically-blind in the sense of distinctive features, but is *sensitive to syllables*, i.e. the alignment is biased towards aligning by syllables. A simple implementation of this with the Needleman-Wunsch affine algorithm was devised. Firstly, the phonetic-blindness was achieved by inputting an identity matrix for the similarity matrix that the algorithm requires. Secondly, the alignment by syllable was done by simply replacing all the vowels with the same segment "V" to represent the nucleus of a syllable, which would therefore bias the alignment algorithm to align by syllables.

Of the four parameters (Match, Mismatch, Gap opening, and Gap extension), the match cost was fixed with the value 1 to minimize the complexity of the optimization; it is also a default value for the match cost in most substitution matrices, therefore only three parameters remain. Manual parameter optimization would be challenging; therefore the Monte Carlo method (Metropolis and Ulam 1949), a computational approach, was employed for finding a suitable set of parameters. The training data was 10% of the corpus which were manually aligned. Half of training data was for calibration and the other half for validation. X (the number of generated sets of parameters) and the upper and lower limits of each parameter were systematically increased until a 100% match rate was achieved.

One disadvantage of this method is that it is not truly phonetically-blind since manual alignment can implicitly introduce phonetic biases. This is not a major drawback in the present study where only unambiguous aligned errors were used. In the future, it might be worth exploring the methodology employed in Hirjee and Brown (2010) where the alignment algorithm was trained iteratively without introducing any phonetic information.

3. Segmental and lexical analyses

The goal of the present analyses is to demonstrate linguistic trends of naturalistic misperception at different levels, starting from the bottom (context-free matrix), and to the top (word frequency). Importantly, this will allow us to identify any significant factors in natu-

ralistic misperception which can subsequently be tested in experiments. The current analyses are based on a corpus of 3,638 naturalistically occurring instances using only Nevins' 2009 and 2010 corpora, and Bond's corpus, since Browman's and Labov's corpora are still in the process of being compiled.

The first step to the analyses was to exclude those instances with 5 or more errors. This reduced the corpus to 2783 instances. The kinds of errors were separated into two kinds, simple segmental errors, e.g. ***p****an* → ***b****an* and complex segmental errors, e.g. ***p****in* → ***sk****in*. To be more specific, simple segmental errors are cases where the immediate adjacent segments of the non-identical segments are identical, e.g. [kʌt] → [kɑt] where [k] and [t] are the same on both intended and perceived. The reason for this separation is to avoid alignment ambiguity in complex segmental errors, given the complexity of deciding which of the following alignments is more valid 1) [**∅p**] → [**sk**] or 2) [**p∅**] → [**sk**] (deletions/insertions are aligned with ∅). The initial analysis was performed with just the simple segmental errors, and focused only on consonant-to-consonant errors.

An overview of consonant errors showed that there were 2712 changed pairs of segments, of which 1598 (59%) were substitutions, 536 (20%) were insertions, 578 (21%) were deletions. Context-free analyses of differential substitution by place were investigated before proceeding to context-sensitive analysis.

3.1 Context-Free segmental analyses: Naturalistic versus Laboratory

The 1598 consonant confusions were tabulated in a confusion matrix, where the diagonal cells were the number of times each segment was correctly perceived in the corpus. The consonant confusions obtained from naturalistic data provided a testbed for comparison with the laboratory perception studies, such as Miller and Nicely (1955), Wang and Bilger (1973) and Cutler et al. (2004).

Many asymmetric confusions with nonsense stimuli were found in the naturalistic errors. Two such prominent pairs, which are also well-known changes in historical and dialectal variation, were [θ]-[f] and [ŋ]-[n]. The naturalistic errors converged with laboratory studies in that [θ]-[f] are consistently confused in the direction of [θ]→[f], for instance, 1) in Miller and Nicely (1955), the CV stimuli set with SNR of +12db and with the frequency bands of 200-5000Hz showed [θ]→[f] at 37% of the time for all the instances of the intended [θ] while [f]→[θ] was 14%; 2) in Wang and Bilger (1973), the CV stimuli set summed over all vowels, noise levels and SNRs again showed [θ]→[f] at 39% of the time while [f]→[θ] was 9.9%; 3) in Cutler et al. (2004), the CV & VC stimuli set summed over all SNRs showed that for the CV set, [θ]→[f] (14%) and [f]→[θ] (9.5%) and for the VC set, [θ]→[f] (36.7%) and [f]→[θ] (13.2%). In the naturalistic errors, the same asymmetry was found with [θ]→[f] at 3.1% and [f]→[θ] at 1.6%.

Again, in laboratory studies, [ŋ]-[n] were consistently confused in the direction of [ŋ]→[n] as found in Cutler et al. (2004), where the VC stimuli set showed [ŋ]→[n] at 17.6% and [n]→[ŋ] at 6.7%. The naturalistic errors also showed this very pattern, [ŋ]→[n] at 5.1% and [n]→[ŋ] at 0.69%.

These similarities suggest that despite all the higher factors on top of phonetics (pragmatics, lexical frequencies and others), the bottom-up phonetic information remains dom-

inant, and these particular confusion pairs provide further evidence that the ongoing sound changes of ŋ-alveolarization and θ-fronting in English are perceptually motivated, lending support to the view of Ohala (1981) that listeners are a source of sound change.

3.2 Featural analyses of place

To explore further than the segmental level, a featural level analysis was performed by looking at the place of articulation (Labial, Coronal and Dorsal). Substitution matrices were analyzed to test the hypothesis of the underspecification of coronal.

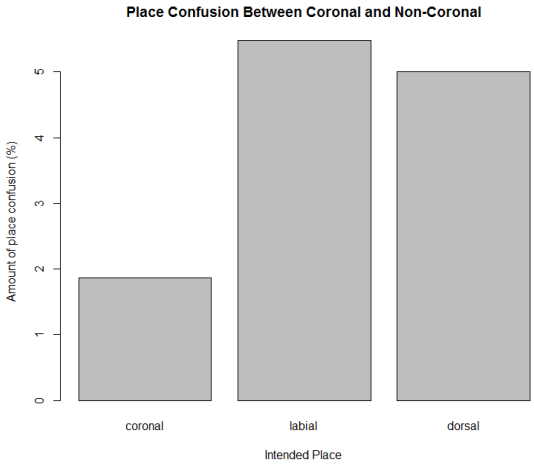
The Featurally Underspecified Lexicon (FUL) model (Lahiri and Reetz 2002) assumes that not all structured features are specified in the phonological representations of morphemes. Under this model of speech perception, listeners compare an incoming speech signal with the set of features in the phonological representation, with either match, mismatch or no-mismatch as outputs. For there to be a *match*, the signal and the lexicon must share the same feature. A *mismatch* requires the signal and the lexicon not sharing the same feature. Finally a *no-mismatch* can happen in several conditions, but the one that is of current interest is when the extracted feature from the signal is underspecified in the lexicon. Under this model, the coronal feature is underspecified, and therefore a hypothesis could be made such that: Pr(Dorsal/Labial $\rightarrow$ Coronal) $>$ Pr(Coronal $\rightarrow$ Dorsal/Labial). This means that the probability of a dorsal or labial segment misperceived as a coronal segment should be higher than the probability of a coronal segment misperceived as a dorsal or labial segment. This is motivated by the model as there is a *no-mismatch* between Dorsal/Labial (acoustic signal) and Coronal (lexicon) which is underspecified; however, there is a *mismatch* between Coronal (acoustic signal) and Dorsal/Labial (lexicon). The no-mismatch condition could contribute to more misperceptions into coronal.

A logistic regression was employed to test the hypothesis that coronals were more often the target than undergoer of misperception. All the aligned pairs of segments was extracted as the sample. If there is a difference between the place of the input segment and the perceived segment, then that pair of segments will be marked as "one" as an attested mistransfer of place, as the comparison here only concerns those confusions between Dorsal/Labial and Coronal (confusions between Labial and Dorsal were excluded, leaving only confusions of Coronal $\rightarrow$ Dorsal/Labial and Dorsal/Labial $\rightarrow$ Coronal).

The function *lmer* in the *R* statistical packages was used. Dummy coding, which compares each level of the categorical variable to a fixed reference level, was used to test for two contrasts: 1) Coronal versus Labial and 2) Coronal versus Dorsal, by setting Coronal as the reference group. In the logistic regression model, the predictors were the place of the input segment as the main effect and the input segments as the random effects, and the predictee was the attested mistransfer of place (one or zero). Figure 2a shows a plot of the distribution between Coronal $\rightarrow$ Dorsal/Labial and Dorsal/Labial $\rightarrow$ Coronal which shows that the Dorsal/Labial $\rightarrow$ Coronal is at least twice as frequent as the other direction. This difference turned out to be statistically significant with Coronal opposed to Labial and Dorsal significantly with $p < 0.01$ for both contrasts; see Figure 2b for detailed output statistics. This result further supports the hypothesis of the underspecification of coronal.

Figure 2: Analyses: Place confusion between coronal and non-coronal

(a) Percentage of place confusion (b) *R* output of the logistic regression model



	β	SE(β)	z	p-value
(Intercept)	-4.0443	.1782	-22.625	$<$.001
Coronal vs Labial	1.169	.3745	3.123	$<$.01
Coronal vs Dorsal	1.125	.4220	2.665	$<$.01

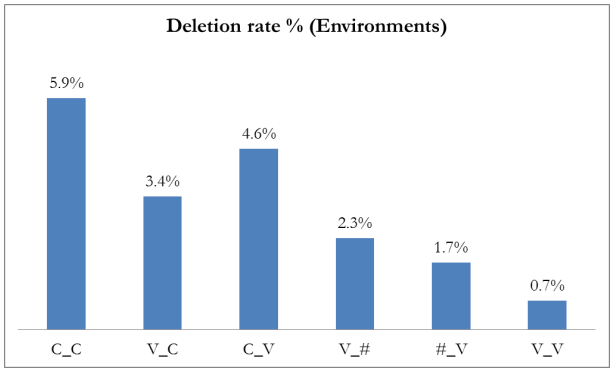
3.3 Adjacency

To explore the data in a context-sensitive analysis analysis looking specifically where certain kinds of deletions and substitutions occur, the effect of segmental adjacency was examined. Environments such as C_C, V_C, C_V, V_#, #_V and V_V were tested for deletions and substitutions. The analyses included both tautosyllabic and heterosyllabic instances of these contexts, inspecting the role of adjacency alone. The hypothesis under investigation is whether the environments with more phonetic cues would be less likely to undergo confusions, and thereby ordering the environments in a hierarchy such as C_C $\succ$ V_C $\succ$ C_V $\succ$ V_# $\succ$ #_V $\succ$ V_V, (where $\succ$ means "has more confusions than"), adapting from Escure (1977)'s environmental hierarchy, and measuring their relative rates of confusion.

Figure 3 shows how the deletion rate varies with each environment. It is clear that environment has a definite effect such that the more phonetically robust the environment is, the lower the deletion rate will be. For example, the interconsonantal segments underwent the most deletions while the intervocalic segments underwent the least deletions. These results appear to fit the hypothesis very well. Similarly, while the substitution rate varies with each environment, the result was contrary to what was expected. While environment still had a definite effect, the effect was in the opposite direction, e.g. intervocalic segments were the most substitutable, while the interconsonantal segments were the least substitutable. Before considering what the possible causes might be, statistical tests were performed to test if these trends were indeed significant.

The significance of adjacency on deletions and substitutions was tested using the function *lmer* in the *R* statistical packages. The contrast was created using Helmert coding (forward and reversed) of the environment, C_C $\succ$ V_C $\succ$ C_V $\succ$ V_# $\succ$ #_V $\succ$ V_V, by comparing each level with the mean of the subsequent(forward)/previous(reversed) levels. The logistic regression model was performed on all the aligned pairs of segments, predicting the attested confusions (one or zero) with the predictors, the six levels of environments being the main effect and the intended segments being the random effects. It was found that out of five contrasts (six levels of environments), at least three were significant for

Figure 3: Analyses: The effect of adjacency on deletion



both deletion and substitution in either coding directions, with the range of p-values: .001 $< p <$ .05. The analyses suggest that the adjacency trend by phonetic cues is on the whole significant.

The apparent asymmetry between deletions and substitutions could be explained in a few ways. One explanation is that in phonetically less robust environments (e.g. C_C), segments get completely deleted, whereas in more robust environments (e.g. V_V), the fact that a segment was there must be retained, as listeners have implicit knowledge that there was a segment to be recovered but are unsure of its identity. As a result, errors, when they occur, are more likely ones of substitution. An alternative explanation would be to consider the phonotactics of English. It may be that the set of possible substitutions is more restrictive than those of deletions. For example, the set of possible substitutions in C_C, V_C, and C_V is more constrained, as the environment f_V is restricted to glides and liquids. On the other hand, in #_V, V_#, and V_V the set of possible substitutions is much larger, thereby increasing the overall rate of substitutability in these environments. This hypothesis awaits further verification by specific segment types.

3.4 Frequency

The effect of adjacency on deletion and substitution appeared to be asymmetrical and complementary. Most of the levels of contrasts were statistically significant. To begin to examine higher level effects in misperceptions, the role of word frequency was analyzed.

The role of frequency, neighborhood density and neighborhood frequency in speech misperception was examined in Vitevitch (2002) using a subset of Bond (1999)'s corpus (88 word pairs) as well as inducing errors experimentally. In Vitevitch (2002)'s comparison of the actual utterance to the perceived utterance, a number of dependent variables were tested, including number of syllables, number of phonemes, familiarity rating, word frequency, neighborhood density, and neighborhood frequency. While one would predict that the perceived words would have higher word frequency, higher neighborhood density, and lower neighborhood frequency than the intended word, it was found that there were no significant differences for all of the dependent measures between the actual and the perceived utterances. Given that the size of the sample used is only 88, this finding would benefit from using a much larger sample, such as the combined corpus in the current paper.

In speech misproduction, Hotopf (1980) examined the role of word frequency using

about 200 naturalistic instances, focusing mainly on semantic slips. Two hypotheses were examined. Hypothesis 1 is that in certain situations, a more frequent word is more likely to be spoken than a less frequent one (e.g. when involving proper names), such that more frequently employed words are more readily accessed than the target word. A plausible hypothesis would thus be for the pronounced word to be of higher frequency than the intended word, Frequency(Perceived) $\succ$ Frequency(Intended). ($\succ$ means "higher than"). Hypothesis 2, on the other hand, would hold that constant and equally frequent use of the names in similar situations has caused them to lose some of their distinctiveness (e.g. calling one's wife by the name of one's daughter). The hypothesis would be for the pronounced word and intended word to be of similar frequency, Frequency(Perceived) $\approx$ Frequency(Intended). In order to test these hypotheses for the naturalistic slips, SUBTLEX-US, a 51 million word frequency corpus (Brysbaert and New 2009), was used. SUBTLEX corpora use film subtitles to construct frequency corpora (see Tang 2012, for example) and these SUBTLEX film subtitle frequencies have been proven to be excellent predictors of behavioral task measures, e.g. English (Brysbaert and New 2009), Dutch (Keuleers et al. 2010) and others. In our combined corpus, 2171 pairs of intended and perceived words were extracted after removing those with zero frequency and duplicates.

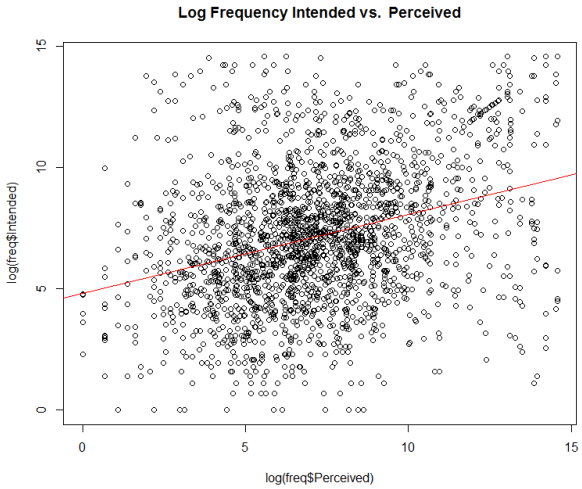
Under Hypothesis 1, it was expected that the log frequency of perceived words should be higher than those of intended words, Frequency(Perceived) $\succ$ Frequency(Intended). Out of the 2171 word pairs, the number of pairs with Frequency(Perceived) $\succ$ Frequency(Intended) is 1072. In the other direction, the number of pairs with Frequency(Intended) $\succ$ Frequency(Perceived) is 1099. A chi-squared test yielded $\chi^2 = 0.3358$, df = 1, p-value = 0.5623, which is statistically insignificant. This suggested that the frequency of perceived words are not significantly higher than the intended, and thus Hypothesis 1 was rejected, a surprising finding.

Under Hypothesis 2, it was expected that the frequency of the perceived word to be similar that of the intended word, Frequency(Perceived) $\approx$ Frequency(Intended). Product moment correlations on the log frequencies of the intended and the perceived word yielded a value of 0.33 (df = 2169, $p < 0.001$, one-tailed), which suggested that the frequencies are significantly correlated, and supported the hypotheses that the perceived and intended words do tend to be of similar frequency (see Figure 4).

The sample of 2171 word pairs thus provides a more convincing account of the role of word frequency in slips of the ear, confirming the direction of Vitevitch (2002)'s findings, with a set of word pairs over 20 times larger. Cutler and Butterfield (1992) also conducted a study with a large number of errors in juncture misperception with 165 instances. They too found no difference between the frequency of actual and perceived utterances. It seems that multiple convergent studies have consistently found similar results on the fact that slips are not biased towards a more frequent perceived-than-intended word, and the consistency across very different studies suggests it is not a statistical artifact.

The finding is in a sense surprising, and several potential explanations are worth exploring. In speech mis*production*, this finding is consistent with Oldfield (1966)'s theory in naming, such that the first choice in the search procedure is of a word-frequency class (Hotopf 1980). However it is not clear how exactly this procedure would directly transfer over to speech *perception*. If the intended word was not recovered, how could the identifi-

Figure 4: Frequency correlation between the intended and perceived words



cation of the word-frequency class be possible for the subsequent retrieval of the perceived word from that class? One explanation was offered by Vitevitch (2002) using the principle of graceful degradation. Graceful degradation is the ability of a processing system to continue operating properly in the event of the failure. In the case of speech misperception, it would translate into the mechanism that when listeners' cognitive systems are faced with incomplete or incorrect information in the speech signal, the "best matches" of the representation of the signal in terms of rough sorting are what is returned (McClelland et al. 1986). Another possibility concerns the relative speech rate of high frequency words, which listeners may know to be pronounced more rapidly (and less clearly) as whole, and thus may surmise that the misperceived word was of a frequency class that corresponds to its usual pronounced speech rate. While the best model of the interaction between frequency and misperception is still not within sight, clearly a simplistic bias towards 'choosing in the direction of words more likely to occur in general' in case of uncertainty cannot be right.

4. Conclusions

The naturalistic data are *consistent* and *complementary* with laboratory results, although much *richer* because they include many more phonological contexts than lab studies (which are usually limited to VCV or CV), and allows analyses beyond the featural level such as environments, word frequency, and potentially further factors such as demographics and conversational topic.

In a preliminary comparison between the naturalistic and experimental data, certain substitutions showed asymmetries in both sets of data, which mirror perceptually-driven phonological processes in English. This suggests that naturalistic data are consistent with laboratory results, although further analyses are needed to confirm this. By examining place confusions, the naturalistic data supported the Featurally Underspecified Lexicon (FUL) model. Furthermore, adjacency environments have clear effects on naturalistic speech perception. The surprising reversed effect of environment of deletions versus substitutions requires further investigation. Finally, the role of frequency was tested with a large set of

data with over 2000 word pairs, and the results supported findings from previous work based on much smaller samples.

Overall, the analyses suggest that phonological and perceptual considerations exert a surprisingly major role in real-life, everyday erroneous performance even when we might expect top-down and contextual effects to otherwise dominate.

References

Bond, Z.S. 1999. *Slips of the ear: Errors in the perception of casual conversation*. New York: Academic Press.

Browman, C.P. 1980. Perceptual processing: Evidence from slips of the ear. In *Errors in linguistic performance: Slips of the tongue, ear, pen and hand*, ed. V.A. Fromkin, 213–230. New York: Academic Press.

Brysbaert, M., and B. New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41:977–990.

Cooke, M., J. Barker, and M.L.G. Lecumberri. in press. Crowdsourcing in speech perception. In *Crowdsourcing in language and speech*, ed. J. Wiley, 137–172.

Cutler, A. 1982. The reliability of speech error data. In *Slips of the tongue and language production*, ed. A. Cutler, 7–28. Amsterdam: Walter de Gruyter/Mouton.

Cutler, A., and S. Butterfield. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language* 31:218–236.

Cutler, A., A. Weber, R. Smits, and N. Cooper. 2004. Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America* 116:3668–3678.

Dell, G.S., and P.A. Reich. 1981. Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior* 20:611–629.

Eidhammer, I., I. Jonassen, and W.R. Taylor. 2004. *Protein Bioinformatics: An algorithmic approach to sequence and structure analysis*. Chichester: Wiley.

Escure, G. 1977. Hierarchies and phonological weakening. *Lingua* 43:55–64.

Ferber, R. 1991. Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue. *Journal of Psycholinguistic Research* 20:105–122.

Fox, C., and R. Combley, ed. 2009. *Longman dictionary of contemporary English, fifth edition (paperback + DVD-ROM)*. Harlow: Pearson Longman.

Hirjee, H., and D.G. Brown. 2010. Solving misheard lyric search queries using a probabilistic model of speech sounds. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 147–152.

Hotopf, W.H.N. 1980. Semantic similarity as a factor in whole-word slips of the tongue. In *Errors in linguistic performance: Slips of the tongue, ear, pen and hand*, ed. V.A. Fromkin, 97–109. New York: Academic Press.

Keuleers, E., M. Brysbaert, and B. New. 2010. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods* 42:643–650.

Kondrak, G. 2003. Phonetic alignment and similarity. *Computers and the Humanities*

37:273–291.
Labov, W. 2010. *Principles of linguistic change, volume III: Cognitive and cultural factors*. Malden, Massachusetts: Wiley-Blackwell.
Lahiri, A., and H. Reetz. 2002. Underspecified recognition. In *Laboratory phonology 7*, ed. C. Gussenhoven and N. Werner, volume 7, 637–676. Berlin: Mouton de Gruyter.
Laver, J. 1970. The production of speech. In *New horizons in linguistics*, ed. J. Lyons, 53–75. Harmondsworth: Penguin.
Lecumberri, M.L.G., A.M. Tóth, Y. Tang, and M. Cooke. 2013. Elicitation and analysis of a corpus of robust noise-induced word misperceptions in Spanish. *Submitted to Interspeech 2013* .
McClelland, J.L., D.E. Rumelhart, and G.E. Hinton. 1986. The appeal of parallel distributed processing. In *Parallel distributed processing: Explorations in the microstructure of cognition*, ed. D.E. Rumelhart, J.L. McClelland, and The PDP Research Group, volume 1, 3–44. Cambridge, MA: MIT Press.
Meringer, R. 1908. *Aus dem Leben der Sprache*. Berlin: B. Behr.
Metropolis, N., and S. Ulam. 1949. The Monte Carlo method. *Journal of the American Statistical Association* 44:335–341.
Miller, G.A., and P.E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America* 27:338–352.
Needleman, S.B., and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48:443–453.
Ohala, J.J. 1981. The listener as a source of sound change. In *Papers from the Parasession on Language and Behavior*, ed. C.S. Masek, R.A. Hendrick, and M.F. Miller, 178–203. Chicago: Chicago Linguistic Society.
Oldfield, R.C. 1966. Things, words and the brain. *The Quarterly Journal of Experimental Psychology* 18:340–353.
Tang, K. 2012. A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics* 24:208–214.
Vitevitch, M.S. 2002. Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear. *Language and Speech* 45:407–434.
Wang, M.D., and R.C. Bilger. 1973. Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America* 54:1248–1266.
Wells, J.C. 2008. *Longman pronunciation dictionary, paper with CD-ROM (third edition)*. Harlow: Pearson Longman.

Division of Psychology and Language Sciences
Department of Linguistics
Chandler House, 2 Wakefield Street
University College London
London WC1N 1PF, United Kingdom

kevin.tang.10@ucl.ac.uk
a.nevins@ucl.ac.uk