Three steps forward for predictability. Consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory.

Introduction

As the studies in this special edition demonstrate, there is now abundant evidence that phonetic forms of linguistic categories (i.e. words and their constituent sounds) are shaped by a number of different factors, and that these governing factors include probabilistic patterns reflecting *predictability* (Shaw and Kawahara 2018). These studies have shown that phonetic forms vary as a function of both phonological properties (e.g. phonotactic predictability) and also message predictability or *informativity* (Cohen Priva 2015, 2017). More controversial, perhaps, are related claims that abstract properties of phonological systems change over time as a consequence of the systematic phonetic patterns shaped by the long-term effects of predictability (e.g. Wedel et al 2013, Bowern & Babinski 2018).

While fully recognising the analytic and theoretical richness of the work on predictability illustrated in this collection, our aim here is to highlight what appears to be a fragility to this work commonly hinging on small differences in acoustic measurements, or interpretations of small statistical effect sizes.

We first highlight a number of caveats about the methods and assumptions encountered in many studies of predictability effects, particularly in reference to corpus-based approaches. Some of the arguments expressed here are relevant not just to predictability work but also to corpus-based research in general. We then consider the wide range of factors that influence patterns of variability in phonetic forms, taking a broad perspective on what is meant by 'the message' in order to show that predictability effects need to be considered alongside many other factors.

We end by suggesting a number of ways forward to extend our understanding of the form-predictability relationship. While some of the points we make are not being made here for the first time, they bear repeating at a time when corpus-based linguistic research represents something of an intellectual gold rush.

Corpus-based methods

Much of the research on predictability effects is characterised by exploitation of large datasets extracted from automatically searchable corpora. (Experimental approaches which control morpho-phonological content or simulate specific speaker or listener effects are increasing, e.g. Lam & Watson 2014, Buz et al 2016, Olejarczuk et al 2018, Tomaschek et al 2018.) Facilitated by the availability of large collections of conversational speech and by the development of increasingly sophisticated automatic analysis methods, corpus-based work is growing rapidly, with applications to many fields including sound change (e.g. Hay & Foulkes 2016) and forensic speaker comparison (e.g. Hughes 2014).

Corpora offer many advantages, reviewed in detail by Harrington (2010). In particular, they provide the scale that enables researchers to tackle questions that can only be addressed by analysing large datasets, such as determining the statistical properties of speech as experienced by members of a speech community. However, corpus-based studies are also characterised by various challenges. Some of these are well known, but others are less widely recognised and their potential effects on

descriptive data and theoretical claims are less clear. This is especially so when the data presented are sets of acoustic measurements. For example, studies of predictability effects have focussed particularly on measures such as VOT (voice onset time), the two lowest formants of vowels, and word durations. The effects reported and interpreted are thus often based on inherently fine-grained and fleeting acoustic details, or phonetic units that are short and problematic to measure even in studio-quality acoustic recordings. Here we highlight four specific issues: (i) within- and between-corpus variability, (ii) labelling of corpora, (iii) resolution of measurements, and (iv) robustness of statistical models.

Within- and between-corpus variability

Many researchers testing hypotheses about the effects of predictability on phonetic variation exploit pre-existing corpora, i.e. resources that have been assembled for a range of different purposes by other research teams using a variety of methods and corpus building and mining tools. For example, the speech samples available might include various modes of talk (interview, conversation, text reading), and varying methods of transmission and recording, in turn yielding different audio qualities. By way of illustration, the ONZE corpus (*Origins of New Zealand English*; Gordon et al. 2007) contains material from speakers born across a range of more than 120 years, and recorded from 1946 to the present day. The technical quality of recordings naturally varies considerably, due to both the recording medium (from acetate discs to digital facilities) and environment (open field to sound-proofed booth).

Technical differences create wide variation in acoustic quality, which in turn is bound to affect any acoustic measurements to be taken and the technical ease with which they can be extracted. Telephone transmission, for instance, removes or damps frequencies below around 300 Hz and above around 3500 Hz, depending on the telephony system. Spectral material outside the 300-3500 Hz range is therefore impaired or totally unavailable for acoustic analysis. This leads in particular to loss of the high frequency energy characterising many fricatives, and to artificial overestimation of the first formant (F1) of vowels, especially with close vowels such as [i] (where F1 naturally falls close to 300 Hz for adult males). Künzel (2001) reports an upshift in F1 of up to 14% in landline recordings. The effects on mobile/cell lines are both stronger and more variable: Byrne & Foulkes (2004) found an average upshift of 29% for F1 compared to simultaneous clean recordings, with all vowel categories significantly affected to some extent. The effects of technical issues on data derived from a corpus are further investigated by De Decker (2016) and Rathcke et al. (2017).

Labelling of corpora

Parsing, labelling and searching of corpora can be done via a range of tools, with or without manual correction of the labelling (Harrington 2010). Forced alignment tools (e.g. LaBB-CAT, MAUS, FAVE) are widely available and improving in accuracy as the years go by, but are still far from perfect. The choice of tools and the extent of manual correction applied to a corpus naturally affect the accuracy of any extracted data.

Automatic labelling faces particular difficulties with recordings that are suboptimal in quality or which involve multiple talkers in natural conversation (Das et al. 2010). Forced alignment, which in principle is usually segment-based, can struggle to identify boundaries where acoustic cues leak across segments. This may potentially occur over long sequences, as e.g. West (1999) shows for cues

to /l/ and /r/. Moreover, natural speech differs markedly from the idealised citation forms typically elicited in the laboratory. The speech signal is more appropriately approached as a series of acoustic landmarks rather than a messy series of segments (cf. Hockett's much cited analogy of Easter eggs on a conveyor belt, smashed by rollers, which captures the massiveness of some surface phonetic variation, but does not recognize its systematicity). That is, the abrupt spectral change events in the speech signal, known as acoustic landmarks (Stevens 2002), and associated with consonantal closures and releases, as well as acoustic minima associated with glides and maxima associated with vowels, provide an initial analysis structure for a spoken waveform which reflects the underlying CV structure of the words. But these time-specifiable acoustic events do not necessarily define regions over which information about a phonemic segment is available; as noted above, information about the features of a segment can be spread over a much larger region of the speech signal. Even more tellingly, comparisons of the performance of forced alignment systems typically report measurement resolution in segment labelling of between 10 and 25 ms (Brognaux et al 2012, Sonderegger & Keshet 2012, Fromont & Watson 2016).

Measurement resolution

An important but rarely acknowledged issue to bear in mind is that, under normal circumstances, there is no objective 'ground truth' when it comes to acoustic analysis. That is, there is no inscrutably 'correct' frequency value or duration measurement. Acoustic analysis is rarely fully straightforward even on the most carefully controlled and articulated material. For example, Duckworth et al. (2011) compared manual formant measurements by experienced analysts on citation form data. Their comparisons showed considerable variation within and between analysts, especially with open vowels, and with variance increasing from F1 to F2 to F3. The mean absolute difference between analysts reached as high as 200 Hz for F3.

Studies such as these remind us that acoustic analysis should be regarded as yielding *estimates* of the quantitative measures at stake rather than inscrutable facts. Those estimates are inevitably sensitive to the technical quality of the material under analysis, and also to decisions made by the analyst in terms of where and how to measure. For example, the frequency range setting for a spectrogram has a powerful influence on measures of high-frequency sounds associated with fricative consonants, amplitude settings influence e.g. whether voice-bars during stop closures are measureable or not, and decisions about which pitch estimation algorithm to use (or even what F0 range to search) can have a significant effect on the estimated F0 value. For such reasons we suggest that the term 'measurement resolution' is generally preferable to 'measurement error'.

Harrison (2013) conducted a detailed investigation of the resolution of formant measurements using different software systems and settings. Recognising the difficulty of quantifying errors, he used synthetic data, specifying formant values *a priori* to provide targets against which accuracy of measurement could be judged. He also used real speech data where 'ground truth' values were available in the form of carefully calculated vocal tract resonances based on acoustic analysis of a subset of the TIMIT corpus (Deng et al. 2006). The results make for sobering reading. For example, with synthetic data the average measurement error for the first three formants was 13 Hz (p. 138). For real speech the errors were far larger and much more variable. Praat's default formant measuring tool with an LPC order of 10 – perhaps the most widely used (if not explicitly acknowledged) method in phonetic studies – resulted in an average error for the first three formants of 96 Hz or 9% (p. 184). Different formants were differently affected, however, and systematic error patterns were observed with changes to LPC order and other settings. Errors were reduced through

changes in settings for different formants and speakers, though it should be borne in mind that fully automatic data extraction normally uses fixed settings for all speakers and tokens being analysed.

Corpus-based studies, including those of predictability effects, vary in the extent to which extracted data are subjected to manual correction. It is reassuring that many studies on predictability effects report either manual labelling or manual correction of automatically-extracted data (e.g. Schuppler et al 2012, Stuart-Smith et al 2015, Buz et al 2016, Hay & Foulkes 2016. Clopper et al 2018), although this is not always the case in other areas of research (see for example Franco-Pedroso & Gonzalez-Rodriguez 2016, who explore uncorrected data for a forensic study). The importance of manual correction is illustrated by Chodroff and Wilson (2017), who found an average 13 ms RMS difference in VOT of voiceless stops when comparing automated and manual measures. However, manual labelling is usually limited to the phonetic variable in focus; automatically-derived measures are also used within many of these studies. For example, Hay & Foulkes (2016) and Chodroff and Wilson (2017) also included articulation rate as a factor in their models, derived automatically from their corpora.

The difficulty inherent in labelling of corpora and extraction of measurements suggests it is prudent to exercise very careful scrutiny of models that report extremely small duration or frequency differences as evidence for predictability effects. These observations further highlight the importance of establishing the degree of reliability for acoustic measurements in any study, and the potential effects on the statistical models reported.

Robustness of statistical models

Variation or errors in a dataset obviously have consequences for statistical models of the data. It goes without saying that any change to a dataset will affect the reported coefficients and associated statistical values. It is equally obvious that bad data can only yield a bad model. However, our aim here is to highlight the *magnitude* of the effects introduced by adjusting the acoustic dataset upon which a model is based. In order to illustrate the potential effects models from Hughes (2014; see also Hughes & Foulkes 2015) were re-run using data that were manually corrected in various ways. Tables 1-3 summarise the results.

The initial purpose of the analysis in Hughes (2014) was forensic, exploring the potential of different variables to discriminate between individual speakers. Analysis was made of several variables, each of which were then used in paired discrimination tasks. Pairs of speech samples were classified via Bayesian likelihood ratios as ‘same’ or ‘different’ speaker, depending on the distribution of the variable in question. In one experiment tokens of the FACE vowel (/ei/) were extracted from male speakers in the ONZE corpus. The LaBB-CAT system with which ONZE is annotated and searched returned measurements capturing the dynamic movement of the first three vowel formants in 10% intervals across the duration of the vowel. The data shown below reflect F1 at the 20% point, as a measure of the vowel’s peak openness, which is expected to display variation as a function of socio-economic class, age, and vowel duration. It was not possible at the time to inspect the data manually (due to the size of the corpus and because manual checking needs to be done *in situ* in New Zealand, for various reasons). A series of heuristics was therefore applied cumulatively by Hughes (2014) to remove potential errors before running models (cf. similar procedures by Hay et al. 2015). At stage 1, outliers were removed if ±3.29 SDs from the mean (following e.g. Tabachnick and Fidell 1996). At stage 2, tokens were removed if they fell outside the expected range based on published literature (a generous 200-900 Hz to avoid being overly prescriptive). At stage 3, tokens were removed if the

measurement represented an unrealistic jump relative to adjacent measurement points (established pragmatically as > 100 Hz).

Tables 1 and 2 summarise the effects on the size of the corpus and the descriptive statistics at each stage of correction. In order to establish the magnitude of effects on the statistical model, the data available at each stage were fitted with the same generic linear mixed effects model using the final data set (i.e. at stage 3). The model included social class, age, and vowel duration as fixed effects and speaker and word as random effects. Table 3 displays the model fit to the data using the r^2 value for both the fixed and random effects at each stage, as well as the percentage improvement in model fit over the previous iteration of the data.

Table 1: corpus size and effects of correction

		corpus size (N data points)	tokens removed	% attrition from original dataset	% attrition from previous stage
	Original dataset	16,960	-	-	-
1	Removal of outliers (> ±3.29 SDs from mean)	16,610	350	2.1	2.1
2	Removal of values outside 200-900 Hz	14,987	1,623	9.8	9.6
3	Removal of jumps > 100 Hz	10,608	4,379	29.2	25.8

Table 2: descriptive statistics on F1 of FACE (in Hz where appropriate) after each correction stage (% change relative to original dataset)

		mean	% change	sd	% change	range	% change	skew	% change	kurtosis	% change
	Original dataset	609	-	157	-	1817	-	1.47	-	10.6	-
1	Removal of outliers	600	-1.5	131	-16.3	1017	-44.0	0.09	-93.8	3.9	-63.1
2	Removal of values outside 200-900 Hz	596	-2.3	119	-24.0	698	-61.6	-0.28	-119.0	3.1	-71.2
3	Removal of jumps > 100 Hz	594	-2.5	113	-28.2	697	-61.6	-0.22	-114.9	3.1	-71.3

Table 3: model fit (r^2) and % change in model fit after each correction stage using a generic linear mixed effects model (fixed effects = class, age, duration; random effects = speaker, word)

		fixed and random effects (conditional)	
		r^2	% change
	Original dataset	0.2369	-
1	Removal of outliers (> ±3.29 SDs from mean)	0.3574	50.8

2	Removal of values outside 200-900 Hz	0.4281	80.7
3	Removal of jumps > 100 Hz	0.4635	95.6

The three stages of correction each had a marked effect on the dataset and statistics. With respect to removal of tokens (Table 1), stage 3 resulted in over 25% of the available data being lost relative to stage 2. The final corpus had seen almost 30% of the original data removed. The descriptive statistics (Table 2) show by stage 3 a reduction in the mean F1 of 15 Hz (-2.5% compared to the original dataset) and large proportional changes in skew (-114.9%) and kurtosis (-71.3%). Table 3 confirms that the model fit improved, as expected with the removal of doubtful data, but the size of this effect is dramatic – the r^2 value of the fixed and random effects almost doubles from 0.2369 to 0.4635 (i.e. the model fit to the data is twice as good after reduction of the dataset compared with using the original data).

A similar conclusion can be derived from another forensically-oriented study by Zhang et al. (2013), who compare formant measurements for the Mandarin triphthong /iau/ using manual and a range of automated measurements. Although the manual measurements were more accurate and yielded the best models, the authors concluded that the manual measurements were only a marginal improvement and thus did not justify the expense of using human analysts in preference to automatic formant tracking. (In a forensic context, where lives could literally be changed by the outcome of an analysis, we would argue that any improvement to that analysis is worth making.) However, when judged proportionally the manual measurements yielded markedly better models. For example, judging from their Figure 2 manual analysis improved the statistical model by around 30% compared to the baseline automatic (MFCC-based) speaker recognition system.

Summary of issues related to corpus-based studies

Corpus tools generate large datasets, but there are many potential sources of error in the data. An implicit assumption in many studies appears to be that noise in the data can be disregarded if the data set is large enough. However, critical evaluation of the raw material is crucial when (i) the object of study is expected to be subtle, such as small differences in VOT, and found intertwined with the effects of many other factors known to affect phonetic form (including social factors, addressed below), and (ii) theoretical claims are advanced on the basis of those measures, without consideration of the resolution with which the measures can be reliably made. The margins of error reported in studies of forced alignment or comparison of automated and manual measurements suggest that any study should cater for a margin of measurement error in the data. Duration measurements are obviously affected if, say, boundaries are inaccurate by 25 ms, but so too will any spectral analysis of incorrectly labelled phones. Manual correction alleviates such problems and also allows the researcher to estimate the margin of error across the corpus. Thus, to increase the usefulness and reliability of corpus results, and to ensure that replication is possible, manual correction ought to be applied to a subset of the data, and the outcome reported in order to gauge the resolution of the data. So too should the effects on any statistical models if they prove sensitive to the adjusted data.

A further consequence of measurement resolution challenges is that we should be careful to translate statistical models back into interpretable units. It has become commonplace in recent years for phonetics researchers to reify statistical models, presenting and discussing the key coefficients of interest as if they are themselves the object of study. Few studies compare model

predictions with the effects in the raw data[1] (Hay & Foulkes 2016 is an exception), or attempt to establish the real-world equivalents of the model predictions (e.g. in terms of milliseconds or Hertz values). The small effects found in some studies, coupled with the uncertainties over accuracy of corpus-based data, highlight how important it is to do so, since differences that cannot be perceived are unlikely to be significant factors in speech signal transmission. While JNDs (just noticeable differences) vary according to stimulus, frequency differences less than 10-20 Hz and duration differences less than 20 ms are unlikely to be perceptible by human beings in real contexts of language use (Stevens 2000: 228-9; but see e.g. Pardo et al. 2017, who report significant effects derived from smaller differences in experiments on conversational convergence). Finally, in an era when acoustic measurements are often reported to several decimal places, it is worth amplifying the advice of Ladefoged (2003) when it comes to instrumental phonetic analysis: to cater for measurement error, reliable data should only be reported in a suitably rounded form (e.g. duration to the nearest 5 ms).

Understanding 'the message'

Our second point when it comes to understanding predictability effects considers what is meant by 'the message'. Usually the message is conceptualised as a purely abstract set of linguistic symbols and associated referential meaning, with variation investigated at the level of e.g. positional allophones or frequency effects on segment or word form in context. However, variation in speech is ubiquitous, highly complex, systematic, and – crucially – *meaningful* on many levels (e.g. Foulkes & Docherty 2006). Variation in phonetic form also reflects systematic influences of (i) other linguistic dimensions such as prosodic structure, (ii) non-linguistic factors such as speaking rate and speaking situation (e.g. amount of background noise), and (iii) a huge range of biological, learned (social) and external factors. Collectively the latter can be termed *indexical* or *sociophonetic* factors.

Prosodic structure

Prosodic structure (hierarchical groupings/boundaries and prominences, often signalled by systematic variation in the values of acoustic parameters such as duration, F0, amplitude, voice quality etc.) may be one of the most significant contributors to systematic context-governed variation in the surface phonetics of word forms. Moreover, the effects of prosodic constituent structure and prominence can be very large. As a result, a thorough understanding of the role of predictability in shaping the phonetic realisation of a word or sound must take this level of linguistic structure into account. However, prosodic structure is rarely taken into detailed consideration in corpus-based research. Turk and Shattuck-Hufnagel (2007) provide an illustration of the importance of prosodic structure. In a study of American English they reported mean phrase-final lengthening effects in read laboratory speech that ranged from 84 ms to 150 ms longer durations of the rhyme of the phrase-final syllable, with individual tokens sometimes exhibiting duration lengthening of 250 ms compared to the same word produced in phrase-medial position. Similarly, Turk and White (1999) reported duration lengthening of 23% (a mean of approximately 30 ms) for syllables that bear phrase-level prominences (pitch accents), compared to unaccented syllables. These effects are not limited to English: for example, Berkovits (1993) reported up to 250% lengthening for phrase-final fricative consonants in Hebrew. This means that the prosodic structure of a speech sample can have a powerful effect on the duration of segments, syllables, words, and strings of words; as a result,

[1] We record our thanks to Jonathan Harrington for raising this issue.

samples that include a higher rate of pitch-accented syllables, or a higher number of phrase boundaries, will show a greater degree of duration lengthening. Moreover, these effects can be considerably larger than the effects often reported for predictability, suggesting the importance of controlling for these prosodic factors when calculating the effect of predictability on duration. Variation in duration is furthermore likely to have indirect effects on spectral properties such as vowel formants, since longer vowels can provide the time required to more closely approximate their canonical formant values (e.g Lindblom 1963).

Indexical factors

Indexical factors convey information about a wide range of potential differences among speakers and speaking contexts, including:

- regional & social background (e.g. age, class, gender, ethnicity, communities of practice);
- speech style (e.g. degree of formality);
- pragmatic intent;
- conversational structure (e.g. cues to turn-taking);
- characteristics of the individual voice (e.g. short-term effects of health, affect & emotion);
- external phenomena (e.g. transmission medium, environmental setting).

Indexical variation is meaningful in that speakers adjust subtle aspects of their speech to convey those indexical properties that are under their control. Listeners in turn respond to the acoustic variation produced voluntarily and involuntarily by speakers, which enables them (among other things) to identify interlocutors, interpret their background, understand their attitudes and emotions, and negotiate spoken interaction. Furthermore, the effects of indexical variation are not separate from the transmission of 'purely linguistic' information. Perceptual processing is affected by systematic variation: for example, we understand words and sounds faster if spoken in a familiar voice (e.g. Nygaard et al. 1994) or a voice with indexical features congruent to the context (e.g. Johnson et al 1999, Walker & Hay 2011). We accommodate to our interlocutors (Bell 1984, Lindblom 1990), for example by drawing on past experience with a speaker or accent when interpreting newly encountered speech (Kleinschmidt & Jaeger 2015), and adjust speech in line with our attitudes towards the listener (Babel 2012) and our perceived success in communication (Buz et al. 2016, Pardo et al. 2017).

When we consider phenomena like phonetic reduction in VOT or shifts in vowel formant patterns as a function of predictability, it is essential to consider both indexical and prosodic factors. The neat conception of full acoustic form = higher information content (and vice versa) is not always tenable. Reduction patterns may themselves be highly informative, perhaps not about the identity of the carrier word but certainly about a speaker's attitude or the ongoing negotiation of a conversation (see also Tomaschek et al 2018). Hawkins (2003) discusses the example of *I don't know* being reduced to [ə̰ə̃ə̰], signalling a great deal about the informality of the situation in which the form was used, and the speaker's attitude to the question previously posed, but without apparently affecting or obscuring referential meaning in context. Local & Walker (2012) discuss a number of subtle but systematic phonetic variations that serve to structure conversational interaction. For example, talk projection (holding the floor in conversation) is cued by articulatory anticipation, avoidance of

lengthening, continued voicing, and segmental reduction, while potential turn transitions are cued by the opposite phonetic effects, including full plosive release.

Those studies that have investigated predictability effects alongside indexical effects invariably find that the latter account for a great deal of the variation in the data, and often considerably more than the effects of predictability factors such as repetition. For example, Hay and Foulkes (2016) conducted a detailed analysis of phonetic, social, lexical and discourse effects on intervocalic /t/ in New Zealand English, which is undergoing a voicing/lenition change from [t] to [d/ɾ]. The lexical and discourse factors included predictability measures related to lexical frequency and whether a word was repeated (i.e. primed) in discourse. Note that the frequency calculation here is based on the whole lexicon contained in the ONZE corpus. It does not take into account segment frequency in context, relative to other segments, and is thus a weaker estimate of predictability that that in e.g. Cohen Priva (2015, 2017). Prosodic effects were not considered in detail, although all tokens were in the same word-internal environment in trochees, e.g. *water*. Table 4 presents the logistic mixed effects regression model for 76 speakers in the ONZE corpus, born from 1932-1982 (based on Table 5 in Hay & Foulkes 2016). The significant effects, calculated as the log odds of [d/ɾ], are presented in decreasing order of size. Social effects (shaded) dominate the model, led by sex and social class. The next two effects relate to discourse topic (discussion of recent events, here in interaction with repetition) and 'word age', a measure of whether a word is used more by younger than older speakers. The predictability effects, as main effects or in interaction, are significant but relatively small (see also Schleef & Turton 2016).

Table 4: regression model for /t/ > [d/ɾ] in New Zealand English, ranked by coefficient size (adapted from Hay & Foulkes 2016, Table 5). Estimates reflect the change in log odds of the variant [d/ɾ].

factor	estimate	SE	z	p	factor type
(Intercept)	-3.517	0.661	−5.317	< 0.0001	
sex = male	2.556	0.356	7.188	< 0.0001	social
class = professional	-1.379	0.347	−3.978	< 0.0001	social
time = recent x repeated = true	1.102	0.365	3.018	< 0.005	social/predictability (interaction)
word age	0.809	0.376	2.153	< 0.05	social
log frequency	0.796	0.151	5.284	< 0.0001	predictability
log frequency x repeated	-0.628	0.207	−3.039	< 0.005	predictability (interaction)
year of birth	0.403	0.179	2.248	< 0.05	social
speech rate	0.291	0.061	4.784	< 0.0001	phonetic
time = recent	-0.120	0.206	−0.585	0.56	social/discourse
repeated	-0.103	0.279	−0.370	0.71	predictability

The statistical model that generated Table 4 was re-run to assess the impact of focusing solely on predictability effects, i.e. ignoring social and other factors. The results of this reduced model are shown in Table 5, and comparison between the two models is summarised in Table 6. Reassuringly, Table 5 shows that the coefficients relating to frequency and the interaction of frequency and repetition remain similar to those shown in Table 4. However, the main effect of repetition now emerges as significant, despite not reaching near to significance in the full model. Table 6 shows

that the full model is a vastly better fit than the reduced model which focuses solely on predictability effects.

Table 5: regression model for predictability effects on /t/ > [d/ɾ] in New Zealand English, ranked by coefficient size (for comparison with Table 4, social effects removed). Estimates reflect the change in log odds of the variant [d/ɾ].

factor	estimate	SE	z	p
(Intercept)	-0.121	-0.2977	-0.406	0.68445
log frequency	0.738	0.1521	4.848	< 0.0001
log frequency x repeated	-0.645	0.2154	-2.993	< 0.01
repeated	0.521	0.1971	2.644	< 0.01

Table 6: comparison based on ANOVA of models shown in Tables 4 (full) and 5 (reduced)

	df	AIC	BIC	logLIK	deviance	Chisq	DF	p
full	13	1722.1	1793.9	-848.06	1696.1	89.724	7	2.2 e-16
reduced	6	1797.8	1831.0	-892.92	1785.8			

Summary

In assessing the results of corpus-based work, we should not assume that several thousand examples of a given segment are all equivalent to each other, shaped only by contextual predictability and linguistic and phonetic factors such as phonotactic position and articulation rate. Instead, an adequate model of predictability effects also needs to consider indexical, stylistic and prosodic factors on an equal footing – these can be at least as important as 'pure linguistic' processing for speaker-listeners in ordinary interaction, and potentially substantially larger.

Ways forward

The discussion above points to a number of areas where the findings of corpus-oriented research into the role of predictability require cautious interpretation (a point echoed by Clopper et al 2018, Daland & Zuraw 2018). In concluding, we suggest three ways in which these analytic and interpretative risks could be mitigated.

First, corpus studies should routinely include a methodological section on the details of corpus tools, data extraction, data analysis, and correction. Explicit consideration should be given to the robustness of models taking into account the measurement resolution of the data. These details are essential for replication of any study, yet are rarely provided. Ideally such studies should also include manual correction of at least a subset of data, in order to estimate the resolution (i.e. the margin of error) in automatic measurements and the relation of this value to the size of effects reported. In addition, it might be instructive to run multiple statistical models with different datasets, corrected in various ways, with a spectrum of results reported where significant variation is obtained. The key findings of statistical analysis should be translated back into real-world phonetic terms and

quantitative units in order to evaluate the robustness of the findings relative to the resolution of the data.

Second, models of data should consider, where possible, the effects of prosodic, stylistic and indexical factors. Sociolinguistically-labelled and prosodically-labelled corpora provide the opportunity to explore those effects alongside acoustic ones (e.g. Hay & Foulkes 2016, Schleef & Turton 2016). It would be valuable to replicate predictability studies using such corpora, to test the robustness of theoretical claims about the phonetic effects of predictability.

Finally, it is vital to test the applicability of predictability models developed on corpus-based datasets in more focussed experimental contexts, where the effects of other factors can be controlled. Given adequate insight from prosodic phonology, sociolinguistics or conversation analysis, it should be possible to replicate corpus-based predictability effects experimentally. It is also possible to test claims made in the latter fields within large corpora. The marriage of corpus and laboratory methods offers huge potential for our understanding of how speech is tailored and understood.

Acknowledgments

We record our thanks to Márton Sóskuthy, two anonymous reviewers and the editors for their helpful suggestions.

References

Babel, M. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics* 40 (1). 177-189.

Bell, Allan. 1984. Language style as audience design. Language in Society 13: 145–204.

Berkovits, R. 1993. Progressive utterance-final lengthening in syllables with final fricatives. *Language and Speech* 36(1). 89-98.

Bowern, C., & Babinski, S. 2018. Mergers in Bardi: Contextual Probability and Predictors of Sound Change. *Linguistics Vanguard* 4(S2).

Brognaux, S., S. Roekhaut, T. Drugman and R. Beaufort. 2012. Automatic phone alignment: a comparison between speaker-independent models and models trained on the corpus to align. Proceedings of the 8th International Conference on NLP, JapTAL 2012, 300–311. Kanazawa, Japan.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. 2016. Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language* 89. 68-86.

Byrne, C. & Foulkes, P. 2004. The mobile phone effect on vowel formants. *International Journal of Speech, Language and the Law* 11. 83-102.

Chodroff, E., & Wilson, C. 2017. Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics* 61. 30-47.

Clopper, C. G., Turnbull, R., & Burdin. 2018. Assessing predictability effects in connected read speech. *Linguistics Vanguard* 4(S2).

Cohen Priva, U. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6. 243–278.

Cohen Priva, U. 2017. Informativity and the actuation of lenition. *Language* 93(3). 569-597.

Daland, R., & Zuraw, K. 2018. Loci and locality of informational effects on phonetic implementation. *Linguistics Vanguard i*(S2).

Das, R., Izak, J., Yuan, J. and Liberman, M. 2010. Forced alignment under adverse conditions. University of Pennsylvania, CIS Dept. Senior Design Project Report.

De Decker, P. 2016. An evaluation of noise on LPC-based vowel formant estimates: implications for sociolinguistic data collection. Linguistics Vanguard 2(1). 83–101.

Deng, L., Cui, X., Pruvenok, R., Chen, Y., Momen, S., & Alwan, A. 2006. A database of vocal tract resonance trajectories for research in speech processing. *Acoustics, Speech and Signal Processing 2006.* 369-372.

Duckworth, M., McDougall, K., de Jong, G., & Shockey, L. 2011. Improving the consistency of formant measurement. *International Journal of Speech, Language & the Law* 18. 35-51.

Foulkes, P. & Docherty, G.J. 2006. The social life of phonetics and phonology. *Journal of Phonetics* 34. 409-438.

Franco-Pedroso, J. & Gonzalez-Rodriguez, J. 2016. Linguistically-constrained formant-based i-vectors for automatic speaker recognition. *Speech Communication* 76. 61-81.

Fromont, R., & Watson, K. 2016. Factors influencing automatic segmental alignment of sociophonetic corpora. *Corpora* 11. 401-431.

Gordon, E., Maclagan, M., & Hay, J. 2007. The ONZE corpus. In J. Beal, K. Corrigan & H. Mosil (eds.) *Creating and digitizing language corpora* (vol. 1), 82-104. Basingstoke: Palgrave Macmillan.

Harrington, J. 2010. *Phonetic analysis of speech corpora*. Malden, MA: Wiley-Blackwell.

Harrison, P. 2013. *Making accurate formant measurements: an empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurement*. PhD dissertation, University of York.

Hawkins, S. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics* 31. 373-405.

Hay, J.B. & Foulkes, P. 2016. The evolution of medial (-t-) over real and remembered time. *Language* 92. 298-330.

Hay, J.B., Pierrehumbert, J.B., Walker, A.J., & LaShell, P. 2015. Tracking word frequency effects through 130 years of sound change. *Cognition* 139. 83-91.

Hughes, V. 2014. *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. PhD dissertation, University of York.

Hughes, V. & Foulkes, P. 2015. The relevant population in forensic voice comparison: effects of varying delimitations of social class and age. *Speech Communication* 66. 218-230.

Johnson, K., Strand, E. A., & D'Imperio, M. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics* 27. 359-384.

Kleinschmidt, D. F., & Jaeger, T. F. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122. 148-203.

Künzel, H. J. 2001. Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics* 8. 80-99.

Ladefoged, P. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Oxford: Blackwell.

Lam, T. Q. & Watson, D. G. 2014. Repetition reduction: Lexical repetition in the absence of referent repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40. 829-843.

Lindblom, B. 1963. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35(11). 1773-1781.

Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (eds.), *Speech production and speech modelling*, 403–439. Amsterdam: Kluwer.

Local, J., & Walker, G. 2012. How phonetic features project more talk. *Journal of the International Phonetic Association* 42: 255-280.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. 1994. Speech perception as a talker-contingent process. *Psychological Science* 5. 42-46.

Olejarczuk, P., Kapatsinski, V., & Baayen, H. 2018. Distributional learning is error-driven: the role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard* 4(S2).

Pardo, J.S., Urmanche, A., Wilman, S., & Wiener, J. 2017. Phonetic convergence across multiplemeasures andmodel talkers. *Attention, Perception and Psychophysics* 79(2). 637-659.

Rathcke, T., Stuart-Smith, J., Torsney, B., Harrington, J. 2017. The beauty in a beast: Minimising the effects of diverse recording quality on vowel formant measurements in sociophonetic real-time studies. *Speech Communication* 86. 24–41.

Schleef, E., & Turton, D. 2016. Sociophonetic variation of like in British dialects: effects of function, context and predictability. *English Language & Linguistics* (first view).

Schuppler, B., van Dommelen, W. A., Koreman, J., & Ernestus, M. 2012. How linguistic and probabilistic properties of a word affect the realization of its final/t: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics* 40. 595-607.

Shaw, J. A., & Kawahara, S. 2018. Predictability and phonology: past, present & future. *Linguistics Vanguard* 4(S2).

Sonderegger, M., & Keshet, J. 2012. Automatic measurement of voice onset time using discriminative structured prediction. *Journal of the Acoustical Society of America 132*(6). 3965-3979.

Stevens, K. N. 2000. *Acoustic phonetics*. Cambridge, MA: MIT press.

Stevens, K. N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America* 111(4). 1872-1891.

Stuart-Smith, J., Sonderegger, M , Rathcke, T. , Macdonald, R. 2015. The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. Laboratory Phonology 6. 505–549.

Tabachnick, B. G. and Fidell, L. S. 1996. Using multivariate statistics (3rd ed.). New York: Harper Collins.

Tomaschek, F., Tucker, B., Fasiolo, M., & Baayen, H. 2018. Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard* 4(S2).

Turk, A. & Shattuck-Hufnagel, S. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 42. 444-472.

Turk, A and White, L. 1999. Structural influences on accentual lengthening in English. *Journal of Phonetics* 27. 171-206.

Walker, A., & Hay, J. 2011. Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology* 2. 219-237.

West, P. 1999. Perception of distributed coarticulatory properties of English/l/ and /r/. *Journal of Phonetics* 27. 405-426.

Wedel, A., Jackson, S., & Kaplan, A. (2013). Functional load and the lexicon: evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech* 56(3). 395-417.

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. 2013. Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison–female voices. *Speech Communication* 55. 796-813.