Learning lexical classes from variable phonology

Stephanie S Shih
University of Southern California

1 Introduction

In morphophonology, it is common for lexical classes to condition different phonological patterns. One particularly well-studied example of this sort is the behaviour of lexical versus grammatical words—i.e., 'content' versus 'function' words[1]—in English. In English, function words are significantly more likely to reduce in surface pronunciation than content words: this reduction manifests on the surface in several ways, including as reduced vowels, in the deletion of segments, and in reduced temporal duration (see e.g., Kelly 1992). Compare, for example, the function word *to* with its lexical homonym *two* in (1). The function word *to* can occur in both full and reduced forms, whereas the content word *two* is less acceptable in the reduced variant.

(1)

		Full	Reduced
a.	*to cats*	✓[tu]	✓[tə]
b.	*two cats*	✓[tu]	*[tə]

Lexical class-conditioned phonology is a very common phenomenon, and numerous types of lexical classes have been shown to engender differences in phonological behaviour: for example,

- morphosyntactic classes (e.g., Kelly & Bock 1988; Smith 2016);
- etymological strata (e.g., Itô & Mester 1999, 2003, 2009);
- semantic classes (e.g., ideophones: Newman 2001; Dingemanse 2012; Rose 2015; Shih & Inkelas 2016);
- extra-linguistic classes (e.g., gender: Slater & Feinman 1985; Cutler et al. 1990; Cassidy et al. 1999; Wright et al. 2005);
- potentially-arbitrary classes (e.g., DuBois 1985).

Morphophonological analyses generally assume that, for lexical class-conditioned phonology, the relevant lexical classes are already known, having been passed to the phonology by either the morphosyntax, some other element of grammar, or the lexicon. However, given that the relevant lexical classes for phonology can range from not just morphosyntactically-defined categories but also to synchronically arbitrary ones, we cannot escape the crucial question of *how* these lexical classes come about:[2] can they be induced from distinctive surface phonological patterns that they trigger? And if so, what lexical classes are learned? If nearly any feature (or non-feature, in the case of arbitrary categories) can delineate a lexical class relevant for conditioning phonological behaviour, then the number of potential lexical classes that one could have interacting with phonology becomes extremely large. Such is the noted problem in lexically-conditioned phonological theory of "cophonology explosion" (e.g., Inkelas et al. 1996). From a gradual learning view, as speakers encounter more lexical items, the number of possible lexical classes relevant for phonology—as individual items, or as classes containing multiple items—increases at a greater-than-exponential rate, fol-

[*] Acknowledgements to Joan Bresnan, Sharon Inkelas, Keith Johnson, Jen Smith, Kie Zuraw, and audiences at UC Merced, UC Berkeley, UCLA, USC, Stonybrook, Edinburgh, McGill, and LSA 2017 for discussion on this work.

[1] The lexical versus grammatical distinction has many names: *lexical/grammatical*, *content/function*, *open/closed*, etc. To prevent confusion with the notion of a lexical class, I use 'content' and 'function' here to refer to lexical and grammatical classes, respectively.

[2] Note that this remains a question regardless of whether we take the common assumption that all possible lexical classes for language are provided by Universal Grammar: somehow the learner must understand *which* of the numerous lexical classes that need to be provided by UG are relevant to their specific linguistic system.

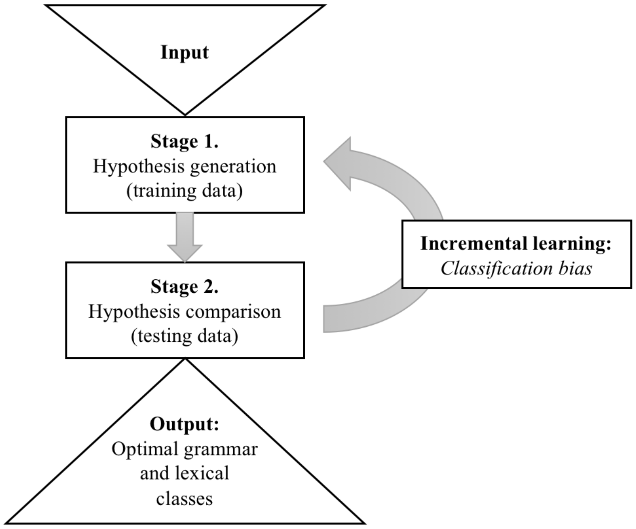
Proceedings of *AJL 2*

lowing the Bell number series (Bell 1934). This problem of discovering relevant classes out of the large range of possibilities is particularly an issue in the face of quantitative variability both within and between lexical classes in natural language, in that we must tease out how much of lexical variability is systematically attributable to lexical class conditioning and how much is simply item-specific variability.

This paper presents a computational approach that induces the relevant lexical classes for lexical class-conditioned phonology from different phonological patterns on the surface, without *a priori* specifications of lexical class. The approach is an unsupervised dual-stage model using bottom-up lexical class discovery and grammar-based hypothesis testing. The current study revisits the lexical class-conditioned phonological behaviours of the supposed categorical distinction between English content and function classes. The classes learned by the computational approach are shown herein to be quantitatively better at capturing lexical class-conditioned phonology, as compared to classes assumed by long-standing, top-down approaches. Given the results, I argue that utilising the dual-stage model with feature discovery and selection within a probabilistic grammar framework allows learners to efficiently hone in on the most robust generalisations of lexical classification over natural language evidence.

1.1 *Model overview* The model proposed here enhances existing approaches to phonological grammar with two crucial components: hypothesis generation and hypothesis testing of probable lexical classifications. The resulting two-stage model of learning lexical classes for the grammar is schematised in (2).

(2) Hypothesis generation and comparison model of lexical class induction



As shown in (2), a learning module first generates a slate of probable, evidence-driven lexical class hypotheses, narrowing down the space of possible lexical classes based on observed surface evidence. Hypothesis generation utilises an agglomerative clustering algorithm. Then, these candidate hypotheses are embedded into the phonological grammar, and the candidate grammars are compared against each other to find the optimal lexical classification, given the weight of evidence. In an incremental implementation, the optimal lexical classification is used as a bias in the subsequent iteration of learning. Endowed with these basic learning and grammar comparison-selection components, observed surface evidence of differential phonological patterns in the input can drive the induction of lexical classes for the grammar in learning.

We can assess the success of lexical class induction using three primary criteria. Lexical classes are likely to be meaningful if they [1] improve the power of the grammar to generate observed data patterns without overfitting; [2] demonstrate independent differences as classes with respect to constraints in the grammar; and [3] concord with existing theoretical expectations and/or previous findings on the expected behaviour of lexical categories. This paper demonstrates the efficacy of the proposed model using a study of phonological reduction conditioned by English content and function classes, as in (1).

The paper is organised as follows. Section 2 introduces the specifics of the data on phonological reduction in English lexical classes. The following sections 3 and 4 illustrate the induction of lexical classes using hypothesis generation and comparison, respectively. Incremental learning is implemented in §5. Section 6 provides a discussion of the results of lexical classification learning, and §7 concludes.

2 Data: Quantifying reduction

The basic observation in the phonological behaviour of content versus function word classes in English is that content words are less likely to reduce than function words. These surface differences between word classes have been argued to reflect deeper differences in stress encoding (e.g., Selkirk 1984, 1996; Kaisse 1985; Inkelas & Zec 1993) or lexical accessibility (e.g., Segalowitz & Lane 2000; Bell et al. 2009). This paper examines the empirical basis of these word class differences: starting from phonological reduction patterns, what are the emergent lexical classes?

The majority view on content and function classes in English (and in other languages) has been one of binarity (e.g., Cann 2000 and references therein): the English lexicon divides neatly into content and function words, and membership in a given class determines the likelihood of reduction. The exact boundaries between content and function classes, however, are not always so clear-cut (see review in Cann 2000), and especially so in their conditioning of phonological reduction. For example, Altenberg (1987) argued that the phonetic correlates of reduction in English did not necessarily point to a simple binary division that is usually taken as the primitive given. Based on a corpus study of spoken English, Altenberg proposed a 10-part classification scale for lexical classes, ranging from groups of words that reduce extremely rarely to those that reduce with great frequency. The middle of the scale includes lexical items such as prepositions and modal auxiliaries that occurred in roughly equivalent reduced and unreduced forms in the corpus study. Subsequent work in speech recognition and corpus linguistics has found similar results on the non-binarity of lexical class-conditioned phonological reduction in English (e.g., Hirschberg 1993; Shih 2014; Anttila 2017).

The phenomenon of content/function-conditioned phonological reduction in English makes for an ideal case study in discovering lexical classes because there is a deep existing literature that outlines clear expectations of how words in content versus function classes should behave in reduction patterns. At the same time, the English case also provides access to sufficient quantitative, natural language data—complete with a good amount of noise and variation—making for a particularly suitable study of whether probabilistic approaches can handle the lexical class induction problem in the face of natural language.

2.1 *Data* The data used here comes from the Buckeye Variation in Conversation corpus (ViC; Pitt et al. 2007). The ViC corpus is a phonetically-transcribed collection of conversational American English from 40 native speakers in Ohio, between 1999–2000. The ViC corpus includes phonetically-transcribed pronunciations, segment and word durations, and part of speech tags based on the Automatic Penn Treebank (Santorini 1990; Marcus et al. 1993). A random subset of the part of speech tags for monosyllabic words in the ViC corpus (*n*=10,000) were hand-checked by the author for accuracy, and if necessary, corrected. The dataset was then annotated for stress, dictionary pronunciation, and syllabic information, taken from the American English Unisyn lexicon (Fitt 2001); words not available in Unisyn were manually coded. Corpus-internal word frequencies, previous and following conditional bigram probabilities, speech rate, and phrase boundaries were calculated via Python scripts.

This study uses all monosyllabic words in the ViC corpus, excluding disfluencies and interjections. The full set of monosyllabic words were trimmed by removing data in which the duration of the word and the average rate of speech surrounding the target word were three standard deviations away from the log mean (following Kuperman & Bresnan 2012). The final dataset includes 206,858 tokens of monosyllabic words, with 2,561 unique orthographic monosyllabic words.

To facilitate reproducibility of lexical classes discussed herein, the Penn Treebank part of speech tags (*N*=39) are used as the basis of determining lexical class membership. While starting with parts of speech can be considered a bit of a shortcut to inducing content versus function classes, it provides a sufficiently complex but manageable level of granularity for considering reduction, versus starting with each terminal lexical item.[3]

2.2 *Quantifying reduction* Previous work has relied largely on listener judgments of prosody and pitch accent when diagnosing stress and reduction, as conditioned by English lexical classes (e.g., Altenberg 1987; Hirschberg 1993). This existing methodology raises at least two issues. First, relying on listener-coded prosodic tagging can lead to inconsistencies, as noted in Ostendorf et al. 2001:2. Hand-labeling also reduces the reliability of and potential for cross-corpus comparisons if different corpora are annotated by different listeners (and usually with different coding schemes). Second, there exists a conceptual gap. On one hand, the empirical corpus work on prosodic differences between function and content words has focused primarily on pitch accent: the presence of a pitch accent is equated with the presence of prominence or stress, broadly construed. On the other hand, the theoretical literature on stress differences between function and content classes focuses primarily on reduction: that is, the presence of reduction is an indicator of the lack of stress. Not only is pitch merely one of several phonetic correlates of stress (Hayes 1995:6), but also the reduction literature usually points to duration and segmental differences (e.g., reduction to schwa)—*not* pitch—as the primary diagnostics of whether reduction has occurred. Some work has been done on reduction from the perspective of segmental alterations (e.g., Johnson 2004a) and duration (e.g. Bell et al. 2003, 2009). The current study follows these in utilising both segmental alterations and duration measures for reduction in a corpus of spontaneous speech.

2.2.1 *Segmental reduction* One phonetic cue to reduction is deviation in segmental material from a full, unreduced form. For example, full vowels reduce to schwa when occurring in unstressed positions. Non-phonologised segmental alterations also occur in reduction in connected speech (e.g. Stampe 1973; Johnson 2004a). Reductions in natural language use can also occur to a certain extent even on stressed syllables: the presence of stress presumably inhibits reduction whereas the absence of stress encourages it. Thus, unstressed syllables should be less faithful to full, lexical forms—that is, more prone to differences in segmental quality and to segmental deletions—than their stressed counterparts.

To quantify segmental reductions, *phone distance*, Johnson 2004a, 2004b's measure of segmental deviation in the ViC corpus, is used here as a proxy of reduction. Phone distance is a calculation that compares the amount of segmental deviation between actual, pronounced phones and their lexical targets: more deviance between the pronounced and target phones is taken as a measure of more phonological reduction. Phone distance is based on the perceptual similarity between phones. Perceptual similarity was calculated by Johnson 2004b via a confusion matrix built on disagreements between transcribers of the ViC corpus. Phone distance for a given word was calculated by summing the perceptual similarity distance between every pair of phones in the pronounced word and the target, citation form of the word. The resulting sum was normalized by Johnson on a scale of 0 to 2, for reasons independent of the study here. All phone distance measures used for the current study were taken directly from Johnson's data.

On the phone distance scale, a phone distance value of 0 represents no deviation between the citation form and the pronounced form. A value of 2 represents the maximum deviation between citation and pronounced forms. For example, when the word *things* is produced as [θɪŋz] in the corpus, it is assigned a phone distance score of 0; when it is massively reduced to [h.ŋs] in production, it is assigned a phone distance score of 1.05. A large portion of the data (44.02%, *n*=91,067) exhibits perfect correspondence between pronounced and citation forms (i.e., phone distance=0), perhaps due to a pressure to produce words faithfully. Very few tokens incur a phone distance value of 2 (*n*=11). All words in this latter set, except one (i.e., *know*), are tokens of *a* and *I*. Seventy-five percent of the data falls below a phone distance value of 0.468.

2.2.2 *Duration* Another cue to reduction is duration. In perception, an increase in syllable duration increases the likelihood the syllable is perceived as stressed, more so than any increase in intensity (Fry 1955). In production, durational differences have been shown to be more closely linked to the differences between stressed and unstressed syllables, whereas differences in pitch denote accentual differences (Beckman & Edwards 1994). The expectation then is that, all else being equal, content words, which are

[3] Even 39 parts of speech provides an incredibly large hypothesis space of lexical classifications: beyond 10 decillion possible lexical class solutions (i.e., B_{39}; Bell 1934).

argued to be lexically stressed and thus lack the ability to reduce, will be longer than function words. For example, *can* as a noun and as a modal auxiliary can be found to be segmentally reduced to [kɪr̃] in ViC. But, these two words—one content word versus one function word—will have differences in durations. As a noun, *can* (produced [kɪr̃]) in *a can of beans* has a duration of 0.295 seconds. In contrast, *can* (also produced [kɪr̃]) as a modal auxiliary in the phrase *I can afford to hang on* exhibits a shorter duration of 0.141 seconds.

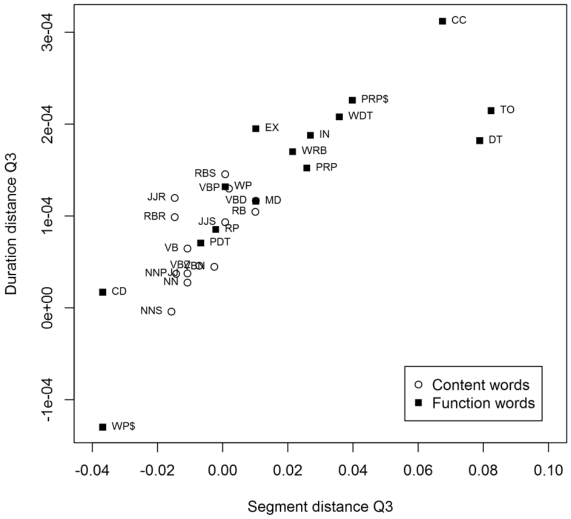
A measure of *duration distance* is used here, as a measure of a produced word's deviation from a target full form. Estimated citation durations of each word were taken by summing, for each phone in the dictionary-listed pronunciation of the word, the average duration of that phone across its every appearance in the corpus. Duration distance was calculated as a ratio of the duration of the produced instance and an estimated citation duration of the word, normalised by the number of phones in both the citation and produced forms (for implementation, see Shih 2014:116ff; see also Gahl et al. 2012 for a similar approach). As with phone distance, a more positive score for duration distance indicates more reduction (i.e., shorter durations than citation forms), and a more negative number indicates less reduction (i.e., longer durations than estimated citation forms). Returning to the example of *can*, the estimated citation duration for the word *can* is 0.284 seconds. The noun instance of *can* cited above (pronounced [kɪr̃] in *a can of beans*) has a duration distance value of -0.00165. In contrast, the modal auxiliary *can* (also pronounced [kɪr̃], in *I can afford to*) incurs a duration distance value of 0.0235. The difference between the two instances of *can* suggests lexical class correlates with durational cues of reduction.

The majority of the data exhibit shorter pronunciation durations as compared to their estimated citation durations (71.64%, *n*=148,198). Because the citation durations used are only estimates, duration distance values smaller than 0 indicate longer durations than expected. These longer-than-expected durations may be due to underestimation in the pronunciation of citation forms when average phone durations were taken from the overall corpus.

2.2.3 *Third quartiles* Lexical classes that exhibit more reduction should have both greater distributional ranges in each of the reduction measures *and* more data mass located at higher values of these measures, indicating higher levels of reduction and greater rates of reduction, respectively. To examine the patterns of phone and duration distance measures together, it is necessary to have a method that captures both of these pieces of information. The standard methods of comparing means and medians do not provide adequate information about ranges *and* data mass at higher ranges. Here, third quartile values of phone distance and duration distance are used instead. The third quartile (Q_3) of a data distribution splits the highest 25% of the data from the lower 75%. Thus, a higher Q_3 indicates more reduction amongst a given set of data, as compared to lower Q_3 values. As opposed to comparing means, comparing third quartiles allows the analysis to capture some information about the upper range of the distribution of reduction measurements in addition to the central mass, particularly when dealing with unknown distributions, as we are here (see, similarity, quantile regression methodologies: Koenker & Bassett, Jr. 1978).

Quartiles for phone distance and duration distance were calculated for each Penn Treebank part of speech tag in the data, using median-unbiased Type 8 quantile estimation in R (Hyndman & Fan 1996). The resulting Q_3 values are plotted against each other in (3), with phone distance represented on the *x*-axis and duration distance on the *y*-axis. For illustration purposes, approximations of content and function word classes are shown, based on previous literature.

(3) Third quartile (Q_3) values of phone distance versus duration distance



3 Generating hypotheses

The first stage of the model of lexical class induction proposed here generates possible hypotheses of likely lexical classes, given observation of phonological reduction patterns, as schematised in (2). A simple clustering algorithm is utilised at this first stage, as a way to narrow down hypotheses, from the vast space of all possible hypotheses. Likely hypotheses are then added to the grammar for comparison and selection in the second stage of the model, §4.

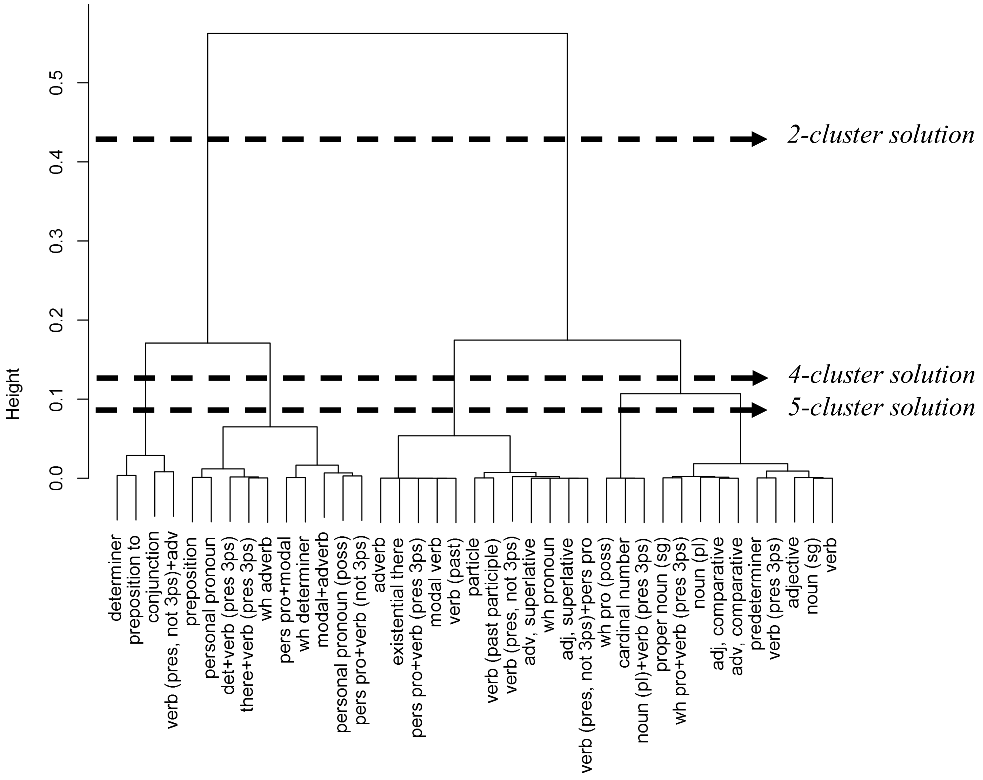
Hierarchical cluster analysis is an unsupervised, agglomerative algorithm for discovering groupings based on similarities and distances between each observation and every other observation on the basis of predictor variables: in this case, between words in each part of speech and every other part of speech group on the basis of reduction (i.e., phone and duration distance). This approach is advantageous for use in the current problem because there is no need to pre-specify the target number of groups that the model should find, as is necessary with many other clustering methods. Because no *a priori* assumptions about the final number of lexical classes are made, this methodology ameliorates preconceived, received-knowledge subjectivity and biases, and allows for a more objective examination of the surface reduction patterns. Hierarchical clustering is thus used here as an approximation of learning from surface data when no lexical class information is yet known.[4] The clustering analysis produces a number of probable lexical class hypotheses based on observed reduction patterns.

Hierarchical clustering is based here on the Q_3 values of phone and duration distance by parts of speech in a training set (80% of data; n=165,486). A test set was held out for the hypothesis comparison and selection component (§4). Phone and duration distance measures were centered and standardised by subtracting their means and dividing by two standard deviations (e.g., Gelman 2008). Q_3 values for the standardised phone and duration distance measures for each part of speech tag were taken. Based on these values, a Euclidean distance matrix was calculated using the `dist` function. The resulting matrix contains the distance between each part of speech tag and every other part of speech tag in the dataset in terms of Q_3 values for the two reduction measures. Ward's method for hierarchical clustering was used (`hclust, method='Ward'`). Ward's method tends to produce equally-sized, tightly-clustered spherical groups, due to its reliance on the sum of squares for within-cluster variance. This means that the clusters found contain members that are all maximally similar in both phone and duration distances.

The results of the hierarchical clustering are given in (4).

[4] That said, the current model does assume that some learning has already happened: e.g., the learner already knows part of speech categories. This is a necessary simplification to scaling this study to a manageable scope.

(4) Hierarchical clustering results: part of speech clusters by phone and duration distances.



A division between content and function classes is clear from the hierarchical model. The right branch of the dendrogram includes parts of speech that are standardly regarded as function words: e.g., determiners, conjunctions, prepositions, *wh* adverbs, and *wh* determiners. The left branch includes not only nouns, verbs, adjectives, and adverbs, but also some grammatical parts of speech: e.g., modals, particles, predeterminers, existential *there*, cardinal numbers (if these are considered 'function' words), and the comparative and superlative adverbs and adjectives. The model also uncovers smaller groups under the largest split based on the phonetic measures of reduction.

The groupings from the hierarchical cluster model can be treated as hypotheses of lexical class solutions: that is, ways in which the lexicon in the training data can be grouped according to the observed surface reduction patterns. These groupings can be seen visually on the dendrogram by tracking horizontal lines across points of the *y*-axis and examining the nodes belonging to each of the vertical lines that intersect the horizontal traces. Two-, 4-, and 5-cluster solutions are illustrated in (4).

4 Selecting hypotheses

Now having narrowed down the space of possible lexical classifications, with observation over the training data, the induction model (2) embeds each lexical class hypothesis into the grammar and compares the performance of these grammars.

4.1 *The grammar* We approximate here a probabilistic 'grammar' of phonological reduction, where the output is a prediction of surface reduction (i.e., phone distance) in previously held-out test data. Constraints used include parameters that are known from the previous literature to affect phonological reduction and to interact with content/function lexical categories. Independent parameters include the following: phrase position, phonological structure (i.e., presence of coda, complex coda, and onset), speech rate, frequency, accent ratio (Nenkova et al. 2007), and conditional probability. For a complete description of the independent parameters used in the model, see Shih 2014:145–148.

The effect of lexical classes on phonological reduction is implemented in the grammar as a main effect and interaction term with the relevant main parameters listed above.[5] This implementation is akin to lexically-indexed cophonologies, or families of lexically-indexed constraints (following Shih & Inkelas 2016; see also Albright 2008; Coetzee & Pater 2011; Moore-Cantwell & Pater 2016). A random effect by lexical item was also included. Grammars were fitted as generalised linear mixed-effects models, using lme4 (Bates et al. 2013). All variables except for multilevel, categorical predictors were centered and standardised. Numerical variables of speech rate, frequency, and conditional probability were log-transformed prior to standardisation.

4.2 *Comparison and selection* We now have several likely hypotheses available about the relevant divisions in the lexical for phonological reduction. The hierarchical cluster analysis in the first stage of the model provides sets of potential lexical classes based on observed reduction by part-of-speech in the training data. These hypotheses range from binary, 2-cluster solutions to possible solutions containing six or more potentially meaningful clusters. From the previous literature, we also have a 'standard', hand-coded content versus function word classification as well as more fine-grained proposals such as Altenberg's (1987) ten-class scale. There is also the null hypothesis that must be considered: that no lexical classification is necessary, and that reduction in speech can alone be predicted by the other parameters in the grammar. In total, 9 lexical class hypotheses are tested here: 6 hypotheses from hypothesis generation, a standard content/function class division, Altenberg's 10-part classification, and the null hypothesis. With hypothesis generation, the vast hypothesis space has now been drastically reduced to a more manageable and reasonable number, from over 10 decillion to about 10.

Each lexical class hypothesis is embedded into the grammar, yielding 9 candidate grammars for comparison. To select the optimal lexical classification, the grammars are compared using information-theoretic model selection, based on Akaike Information Criteria (AIC; Akaike 1973; Anderson & Burnham 2002; Burnham & Anderson 2002, 2004). AIC-based model selection is founded on the idea that all grammars (i.e., models) are mere approximations of full reality, an ideal for which the true parameters cannot be known. Because we are only providing estimations of full reality in our grammars, every grammar experiences some amount of information loss when compared to the "true grammar" of reality: there will also be an amount of uncertainty that does not capture full truth. AIC measures how much information is lost between full reality and the estimated grammar (following Kullback & Leibler 1951). Comparing the differences between AICs of candidate estimated grammars provides a way to quantify the relative strength of evidence in favour of each candidate grammar. In this way, it is possible to compare how well each lexical classification hypothesis produces a parsimonious grammar that suffers the least amount of information loss while at the same time positing the fewest number of necessary parameters (e.g., Shih 2017; see also Wilson & Obdeyn 2009 for similar approach in assessing phonological grammars).

Second order AIC (AIC$_C$) is used, which adjusts for the number of parameters in a model (5):

(5) $$AIC_C = -2log\left(\mathcal{L}(\hat{\beta}|D)\right) + 2K + \frac{2K(K+1)}{n-K-1},$$
where $\mathcal{L}(\hat{\beta}|D)$ = maximum likelihood of observed data D given fitted parameters $\hat{\beta}$,
K = number of estimable parameters (i.e., constraints) in the model,
and n = sample size.

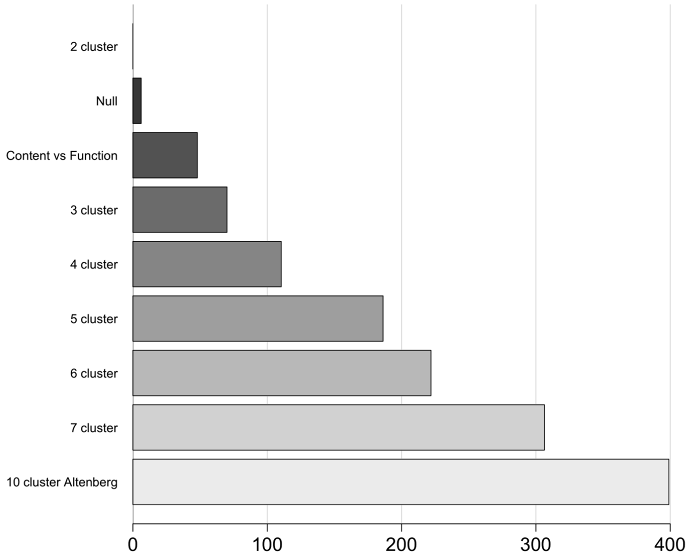
Algebraically lower AIC$_C$ values indicate better candidate models, given the available evidence. AIC$_C$ was calculated using the MuMIn package (Bartoń 2013).

4.3 *Comparison and selection results* The AIC$_C$ results of comparing the 9 candidate grammars are given in (6). Figure (6a) gives AIC$_C$ comparisons over the 20% (*n*=41,370) held-out test data; (6b) gives the AIC$_C$ comparisons over a larger set of randomly sampled data (*n*=124,166).

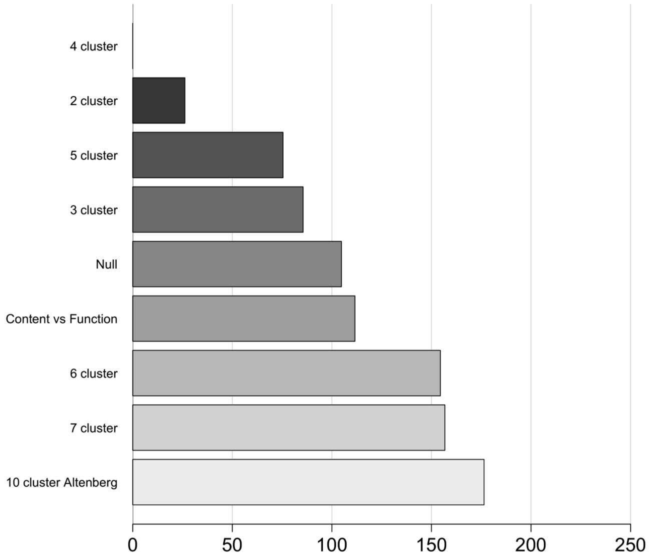
[5] Since it is assumed here that lexical classes fall along a sequential scale for triggering reducibility, reverse Helmert coding for the factor levels was used when there were more than 2 lexical classes.

(6) ΔAIC$_C$ results

a. ΔAIC$_C$ from best model for 20% held-out test sample.



b. ΔAIC$_C$ from best model for 60% random test sample



The results in (6) show that having as many as 10 classes, as hypothesized in Altenberg 1987, is not supported by the evidence: more parameters (i.e., lexical classes) are used than are needed. The top-ranking lexical classifications are those on the lower end of the range, utilising between 2 and 4 lexical classes. As the testing data increases, the need for lexical classes that allow for more category-specific subtlety also increases. For instance, when tested over the smaller set of data (6a), the hypotheses that use fewer lexical classes (2-cluster, null hypothesis, and standard content/function classes) have the lowest AIC$_C$ scores. With more data, hypotheses that have more divisions in the lexicon (e.g., 4-cluster) have lower AIC$_C$ scores, indicating that the increase in parameters and complexity in the model is justified given increases in variance in the data.

In the following section, incremental learning will demonstrate that the optimal classification solutions for English lexical/grammatical prosodification tend to be those using between 3 to 5 lexical classes, such that the relevant lexical classes for reduction mirror standard descriptions of lexical versus grammatical classes, but allow for some category-specific subtlety.

5 Incremental learning

The results in §4 demonstrate that lexical categories can be batch-learned. This section presents an incremental simulation using the model laid out in (2).

In the incremental version of the hypothesis generation and comparison model, each iteration through generation and comparison introduces an unseen 0.1% (*n*=241) of the data for training. Classification hypotheses for each iteration are compared over a new test set of randomly-drawn 20% (*n*=37,235) of the data. The top-ranking hypothesis is chosen, and the class membership of each part of speech is then passed onto the next iteration of the model in the form of a cumulative bias, B.

Bias is implemented as the summation of all current and previous b values, each randomly selected from a Gaussian distribution, (7):

(7) $B = \sum_{k=1}^{i} b_k$,
where $b \leftarrow N(-x, x)$,
and $N(-x, x)$ is a Gaussian distribution over range $(-x, x)$.

In this way, distinct class memberships are not passed from each learning iteration to the next. Instead, as words behave together in selected optimal classifications over time, their cumulative B scores become more similar to each other than words that are not in the same recurring classifications. This implementation of bias thus has the design advantage that classes are not remembered on a 'scale' relative to other classes, but instead are simply remembered as words that behave together, following Hebbian learning (Hebb 1949).

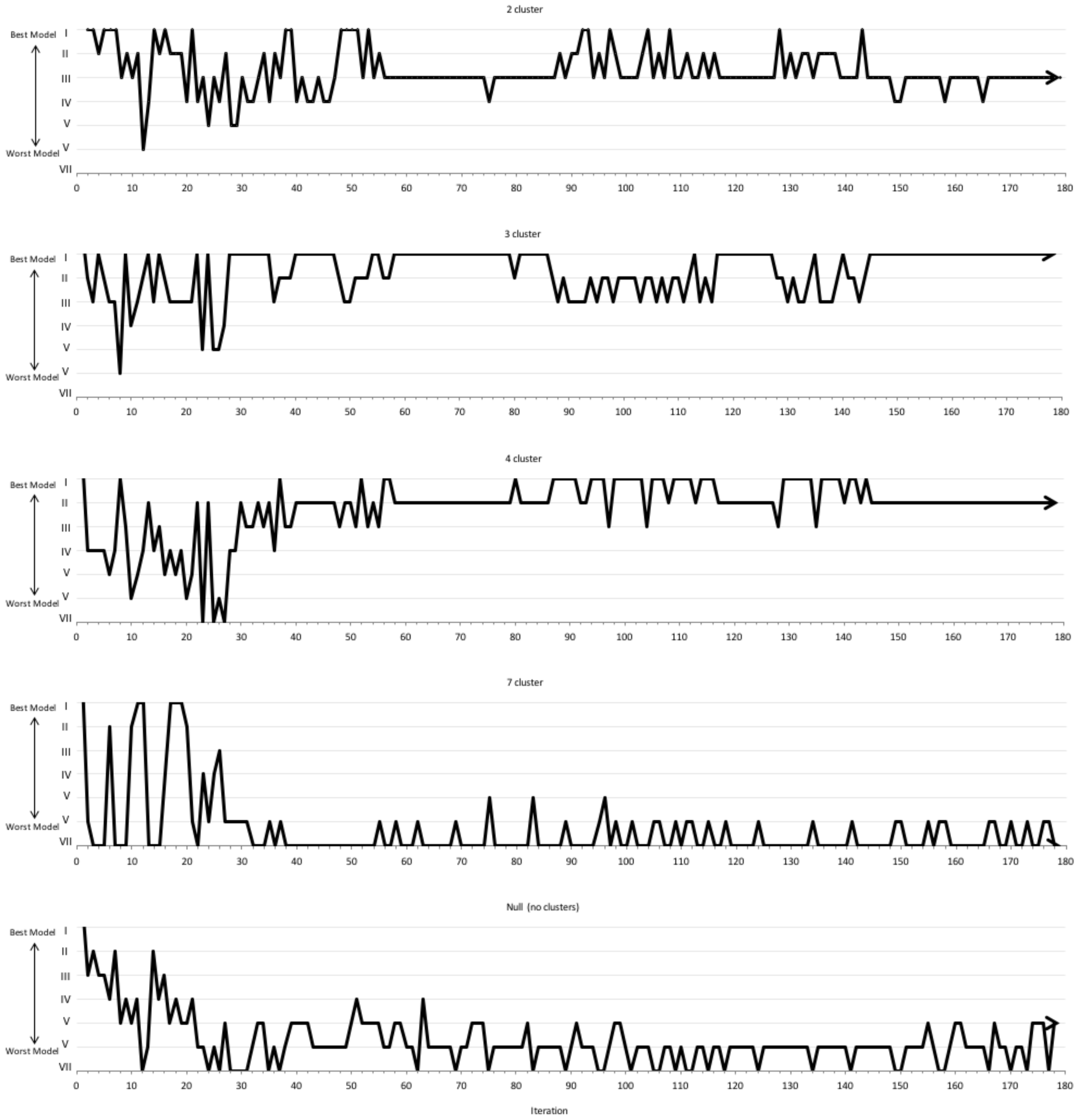
Representative results from an incremental learning simulation are given in (8). The simulation was run over 180 iterations (in B, *x*=0.0001), and each lexical class hypothesis in each iteration is graphed on the *y*-axis with respect to the other hypotheses, from best to worst model in the AIC$_C$ comparison. As the results show, lexical classification rankings start out chaotic while the learner attempts to deal with new input data that may contradict its previously optimal assumptions of lexical classification. For instance, having no clusters (i.e., the *null* condition) is often chosen as the optimal classification solution at the early stages. Over time, 3- and 4-cluster lexical classifications become the most robust optimal hypotheses, reaching relative stability by about 50 iterations. The 2-cluster hypothesis ranks third. Crucially, the iterative learning demonstrates that, over natural language data input, it is not the case that the learning algorithm will justify more and more fine-grained lexical classes when presented with more data. Instead, the learning algorithm proposed here is constrained by a penalty on increasing numbers of parameters during the model comparison component (see 5). As such, the number of lexical classes will increase *only* when there is a sufficient amount of variance in the evidence that demonstrates that additional classes are necessary for the grammar's goodness-of-fit, *not* simply when there is more data presented.

The specification of the normal distribution, $N(-x, x)$, from which b is randomly drawn can be adjusted as desired: larger b values result in reaching robust distinctions between lexical classes in fewer iterations than for smaller b values. It is possible, in the future, to modulate the bias factor in various ways: for example, to fade with memory (see e.g., Moore-Cantwell (2017)) and to incorporate a classification's confidence score at each iteration. The results of the incremental simulation show that classifications can be learned incrementally, and that the optimal number of lexical classes stabilises as the learning progresses.

6 Discussion: an assessment of learned lexical classes

We return here to the assessment of lexical classifications found. Three criteria for successful lexical class induction were laid out at the outset in §1.1. Lexical classes are likely to be meaningful if they [1] improve the power of the grammar to generate observed data patterns without overfitting; [2] demonstrate independent differences as classes with respect to constraints in the grammar; and [3] concord with existing theoretical expectations and/or previous findings on the expected behaviour of lexical categories.

(8) Incremental learning simulation: AIC$_C$ model rankings over 180 iterations



Model comparison has shown that the lexical classifications that are found by the model presented above improve the power of the grammar to generate observed data patterns without over-specifying unnecessary parameters. We turn here to the other two criteria by examining the behaviour of the model with the 4-cluster lexical class hypothesis found from clustering. In particular, we focus on whether the lexical classes found are sensible, given our expectations of how content/function words should behave in reduction patterns.

The table in (9) provides the part-of-speech membership in each cluster of the 4-cluster lexical class hypothesis found during hypothesis generation (see 4).

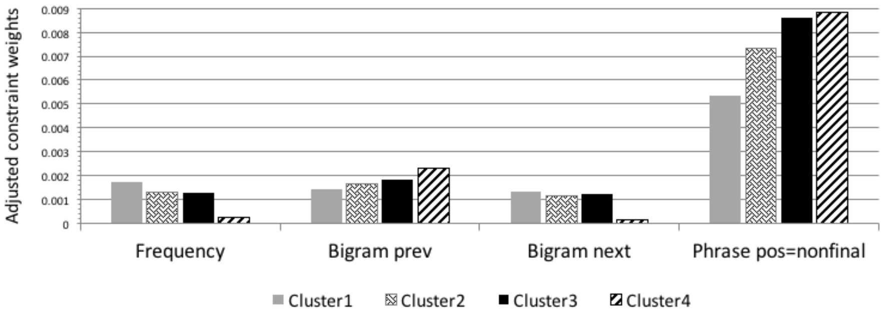
(9) Membership in 4-cluster lexical class hypothesis

Cluster 1 cardinal numbers, possessive *wh* pronoun, adjectives, nouns, proper nouns, predeterminers, present tense verbs (3p), some adverbs
Cluster 2 particles, modal auxiliaries, some adverbs, past tense verbs, present tense verbs (not 3p), *wh* pronoun, existential *there*
Cluster 3 prepositions, pronouns, *wh* adverb, *wh* determiner
Cluster 4 conjunctions, determiners, preposition *to*

When tested in batch learning over 60% randomly sampled data, this 4-cluster hypothesis ranked the best out of all other lexical class solutions; over a 20% held-out test set, this 4-cluster hypothesis ranked fifth, behind 2- and 3-cluster solutions (see §4). In iterated learning, 4-cluster solutions were amongst the first- and second-best hypotheses tested (see §5). The cluster membership in (9) also demonstrates that the words that are sorted into classes are relatively expected: that is, we do not find wildly surprising class memberships. For example, modal auxiliaries, which are traditionally problematic in content/function classification, are placed in the same cluster as some of the inflected verbs, and not with, say, less “content-ful” words like prepositions (which are in Cluster 3). Thus, words are classified alongside the other words regularly share similar reduction patterns.

The figure in (10) provides the constraint weights for each cluster of the 4-cluster lexical class hypothesis (see 4), for a number of constraints in the grammar. A lower constraint weight indicates that the parameter is less important for that particular lexical class.

(10) Lexical class constraint weights for 4-cluster grammar



It has been previously observed that lexical words are more sensitive to lexical frequency-induced reduction (Bradley 1978; Bell et al. 2009; cf. Gordon & Caramazza 1982; Segalowitz & Lane 2000). The results of the grammar utilising the 4-cluster lexical class hypothesis corresponds to this expected behaviour. As demonstrated in (10), the effect of frequency scales by cluster. Cluster 4 words especially exhibit significantly less reduction compared to other clusters.

Another previously observed difference between English content/function classes is that function words are more sensitive to predictability-induced reduction (Bell et al. 2009). We see that this is evident in the lexical classes presented here, as well. Both preceding (i.e., *bigram prev*) and following (i.e., *bigram next*) bigram predictability effects are magnified for Cluster 4 words, compared to other classes. For preceding bigram predictability, this means that there is more reduction of the more function-like words when the preceding bigram predictability increases than for more content-like words. For following bigram predictability, the reverse is true.

Finally, it is known from previous work that function words are more susceptible to phrase-internal reduction than content words (Selkirk 1984, 1996; Kaisse 1985; Inkelas & Zec 1993). The four lexical classes found by the hypothesis and comparison model demonstrate this prediction. All else being equal, phrase-internal reduction becomes significantly more common with each successive lexical class towards the function-like end of the spectrum: Cluster 2 words are more prone to reduction phrase-internally than Cluster 1 words, and Cluster 3 words are more prone to reduction phrase-internally than Cluster 2 and Cluster 1 words, and so on.

By exploring the 4-cluster solution in detail, we see here that the lexical classes learned by the hypothesis generation and comparison model show independent effects in the grammar and concord with our prior expectations of content and function class behaviour. Thus, the three criteria of identifying induction success are met.

7 Conclusion

The case study on content and function word classification presented here has demonstrated that phonologically-relevant lexical categories can be induced from the surface, via a model that incorporates hypothesis generation and information-theoretic hypothesis comparison and selection. Using a basic hypothesis generation algorithm, the search space of possible lexical class hypotheses is cut down significantly to a more efficient and reasonable size. The learner then balances explanatory power against having too much complexity in the grammar, and learns relevant lexical categories only when they prove useful for the data encountered. This model presents an avenue to linguistic feature discovery that relies not on an *a priori* listing of all possible parameters but instead on observation-based learning (see e.g., Hayes & Wilson 2008 for another on-line feature induction algorithm, albeit for constraints, not classes).

The model presented herein is a computational approximation of how learners might narrow down the search space for relevant lexical categories, but of course, other cues—semantic, syntactic, and extra-linguistic—work in conjunction with phonetic and phonological ones during category acquisition. The model here is flexible enough to be able to incorporate such information at either the hypothesis generation or the hypothesis comparison stage: where exactly such non-phonological factors enter the process, we leave to future work. Furthermore, while it has been shown here that hypothesis generation-and-comparison is a computationally viable model of lexical class induction, it remains an open question whether such an approach is a cognitively viable model of human morpho-phonological category learning during acquisition.

References

AKAIKE, HIROTUGU. 1973. Information Theory as an Extension of the Maximum Likelihood Principle. *Second Interational Symposium on Information Theory*, ed. by B.N. Petrov and F. Csaki, 267–281. Budapest: Akademiai Kiado.

ALBRIGHT, ADAM. 2008. How many grammars am I holding up? Discovering differences between word classes. *Proceedings of the 26th West Coast Conference on Formal Linguistics*, ed. by Charles B. Chang and Hannah J. Haynie, 1–20. Somerville, MA: Cascadilla Proceedings Project.

ALTENBERG, BENGT. 1987. *Prosodic patterns in spoken English: Studies in the correlation between prosody and grammar for text-to-speech conversion*. Lund Studies in English 76. Sweden: Lund University Press.

ANDERSON, DAVID R.; and KENNETH P. BURNHAM. 2002. Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management* 66.912–918.

ANTTILA, ARTO. 2017. Stress, Phrasing, and Auxiliary Contraction in English. *The morphosyntax-phonology connection: Locality and directionality at the interface*, ed. by Vera Gribanova and Stephanie S Shih, 143–170. Oxford, UK: Oxford University Press.

BARTOŃ, KAMIL. 2013. *Package "MuMIn."* http://cran.r-project.org/web/packages/MuMIn/index.html.

BATES, DOUGLAS; MARTIN MAECHLER; and BEN BOLKER. 2013. *Package "lme4."* http://cran.r-project.org/web/packages/lme4/lme4.pdf.

BECKMAN, MARY; and J. EDWARDS. 1994. Articulatory evidence for differentiating stress categories. *Phonological structure and phonetic form: Papers in Laboratory Phonology III*, ed. by Patricia Keating, 7–33. Cambridge University Press.

BELL, ALAN; JASON M. BRENIER; MICHELLE GREGORY; CYNTHIA GIRAND; and DANIEL JURAFSKY. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60.92–111.

BELL, ALAN; DANIEL JURAFSKY; ERIC FOSLER-LUSSIER; CYNTHIA GIRAND; MICHELLE GREGORY; and DANIEL GILDEA. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113.1001–1024.

BELL, ERIC TEMPLE. 1934. Exponential polynomials. *Annals of Mathematics* 35.258–277.

BRADLEY, DIANNE CHRISTINE. 1978. *Computational distinctions of vocabulary type*. Cambridge, MA: Massachusetts Institute of Technology Ph.D. dissertation. http://hdl.handle.net/1721.1/16463.

BURNHAM, KENNETH P.; and DAVID R. ANDERSON. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd edition. New York, NY: Springer-Verlag.

BURNHAM, KENNETH P.; and DAVID R. ANDERSON. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* 33.261–304.

CANN, RONNIE. 2000. Functional versus lexical: A cognitive dichotomy. *The Nature and Function of Syntactic Categories*, ed. by Robert D. Borsley, 37–78. Syntax and Semantics 32. San Diego, CA: Academic Press.

CASSIDY, KIMBERLY WRIGHT; MICHAEL H. KELLY; and LEE'AT J. SHARONI. 1999. Inferring gender from name phonology. *Journal of Experimental Psychology: General* 128.362–381.

COETZEE, ANDRIES W; and JOE PATER. 2011. The place of variation in phonological theory. *The Handbook of Phonological Theory*, ed. by John Goldsmith, Jason Riggle, and Alan Yu, 401–431. 2nd Edition. Cambridge: Blackwell.

CUTLER, ANNE; JAMES MCQUEEN; and KEN ROBINSON. 1990. Elizabeth and John: Sound Patterns of Men's and Women's Names. *Journal of Linguistics* 26.471–482.

DINGEMANSE, MARK. 2012. Advances in the Cross-Linguistic Study of Ideophones. *Language and Linguistics Compass* 6.654–672.

DUBOIS, JOHN W. 1985. Incipient semanticization of possessive ablaut in Mayan. *International Journal of American Linguistics* 51.396–398.

FITT, SUSAN. 2001. Unisyn lexicon. Centre for Speech Technology Research, University of Edinburgh. www.cstr.ed.ac.uk/projects/unisyn.

FRY, DENNIS. 1955. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America* 35.765–69.

GAHL, SUSANNE; YAO YAO; and KEITH JOHNSON. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66.789–806.

GELMAN, ANDREW. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27.2865–2873.

GORDON, BARRY; and ALFONSO CARAMAZZA. 1982. Lexical Decision for Open- and Closed-Class Words: Failure to Replicate Differential Frequency Sensitivity. *Brain and Language* 15.143–160.

HAYES, BRUCE. 1995. *Metrical stress theory: Principles and case studies*. Chicago, IL: The University of Chicago Press.

HAYES, BRUCE; and COLIN WILSON. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440.

HEBB, DONALD O. 1949. *The Organization of Behavior*. New York, NY: Wiley & Sons.

HIRSCHBERG, JULIA. 1993. Pitch accent in context: predicting intonational prominence from text. *Artificial Intelligence* 63.305–340.

HYNDMAN, ROB J.; and YANAN FAN. 1996. Sample quantiles in statistical packages. *The American Statistician* 50.361–365.

INKELAS, SHARON; C. ORHAN ORGUN; and CHERYL ZOLL. 1996. *Exceptions and static phonological patterns: cophonologies vs. prespecification*. UC Berkeley and U. Iowa.

INKELAS, SHARON; and DRAGA ZEC. 1993. Auxiliary reduction without empty categories: a prosodic account. *Working Papers of the Cornell Phonetics Laboratory*, 205–254. 8. Cornell University.

ITÔ, JUNKO; and ARMIN MESTER. 1999. The phonological lexicon. *The Handbook of Japanese Linguistics*, ed. by Natsuko Tsujimura, 62–100. Oxford: Blackwell.

ITÔ, JUNKO; and ARMIN MESTER. 2003. *Japanese Morphophonemics: Markedness and Word Structure*. MIT Press Linguistic Inquiry Monograph Series 41. Cambridge, MA: MIT Press.

ITÔ, JUNKO; and ARMIN MESTER. 2009. Lexical classes in phonology. *The Oxford Handbook of Japanese Linguistics*, ed. by Shigeru Miyagawa and Mamoru Saito, 84–106. Oxford, UK: Oxford University Press.

JOHNSON, KEITH. 2004a. Massive reduction in conversational American English. *Spontaneous Speech: Data and Analysis.*, 29–54. Tokyo, Japan: The National International Institute for Japanese Language.

JOHNSON, KEITH. 2004b. Aligning phonetic transcriptions with their citation forms. *Acoustics Research Letters Online* 5.19–24.

KAISSE, ELLEN M. 1985. *Connected Speech: The Interaction of Syntax and Phonology*. Orlando, Florida: Academic Press, Inc.

KELLY, MICHAEL H. 1992. Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review* 99.349–364.

KELLY, MICHAEL H.; and J. KATHRYN BOCK. 1988. Stress in time. *Journal of Experimental Psychology: Human Perception and Performance* 14.389–403.

KOENKER, ROGER; and GILBERT BASSETT, JR. 1978. Regression Quantiles. *Econometrica* 46.33–50.

KULLBACK, SOLOMAN; and RICHARD A. LEIBLER. 1951. On Information and Sufficiency. *Annals of Mathematical Studies* 2.79–86.

KUPERMAN, VICTOR; and JOAN BRESNAN. 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* 66.588–611.

MARCUS, MITCHELL P.; BEATRICE SANTORINI; and MARY ANN MARCINKIEWICZ. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19.313–330.

MOORE-CANTWELL, CLAIRE. 2017. Concurrent learning of the lexicon and phonology. Paper presented at the Symposium: Learning Lexical Specificity in Phonology. Linguistic Society of America 91st Annual Meeting, Austin, TX.

MOORE-CANTWELL, CLAIRE; and JOE PATER. 2016. Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints. *Catalan Journal of Linguistics 15*, ed. by Eulàlia Bonet and Francesc Torres-Tamarit.

NENKOVA, ANI; JASON M. BRENIER; ANUBHA KOTHARI; SASHA CALHOUN; LAURA WHITTON; DAVID BEAVER; and DANIEL JURAFSKY. 2007. To memorize or to predict: Prominence labeling in conversational speech. *Proceedings of NAACL HLT 2007*, 9–16. Rochester, NY.

NEWMAN, PAUL. 2001. Are ideophones really as weird and extra-systematic as linguists make them to be? *Ideophones*, ed. by F. K. Erhard Voeltz and Christa Kilian-Hatz, 251–258. Typological Studies in Language 44. Amsterdam: John Benjamins.

OSTENDORF, MARI; IZHAK SHAFRAN; STEFANIE SHATTUCK-HUFNAGEL; LESLEY CARMICHAEL; and WILLIAM BYRNE. 2001. A prosodically labeled database of spontaneous speech. *ISCA Tutorial and Research Workshop on Prosody and Speech Recognition and Understanding*. Red Bank, NJ: International Speech Communication Association. http://www.isca-speech.org/archive_open/archive_papers/prosody_2001/prsr_022.pdf.

PITT, MARK A.; LAURA DILLEY; KEITH JOHNSON; SCOTT KIESLING; WILLIAM D. RAYMOND; ELIZABETH HUME; and ERIC FOSLER-LUSSIER. 2007. Buckeye Corpus of Conversational Speech (2nd release). Department of Psychology, Ohio State University. www.buckeyecorpus.osu.edu.

ROSE, SHARON. 2015. Phonology of Ideophones in African Languages. Paper. Paper presented at the WOCAL 8, Kyoto University, Japan.

SANTORINI, BEATRICE. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*. Department of Computer & Information Science Technical Reports (CIS). Philadelphia, PA: University of Pennsylvania. http://repository.upenn.edu/cis_reports/570.

SEGALOWITZ, SIDNEY J.; and KORRI C. LANE. 2000. Lexical access of function versus content words. *Brain and Language* 75.376–389.

SELKIRK, ELISABETH. 1984. On the major class features and syllable theory. *Language Sound Structures*, ed. by Mark Aronoff, R. T Oehrle, Mark Aronoff, and R. T Oehrle, 107–136. Cambridge, Mass.: MIT Press.

SELKIRK, ELISABETH. 1996. The prosodic structure of function words. *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, ed. by James L Morgan, Katherine Demuth, James L Morgan, and Katherine Demuth, 187–214. Lawrence Erlbaum Associates.

SHIH, STEPHANIE S. 2014. *Towards Optimal Rhythm.* Stanford, CA: Stanford University PhD dissertation.

SHIH, STEPHANIE S. 2017. Constraint conjunction in weighted probabilistic grammar. *Phonology*.243–268.

SHIH, STEPHANIE S; and SHARON INKELAS. 2016. Morphologically-conditioned tonotactics in multilevel Maximum Entropy grammar. *Proceedings of the 2015 Annual Meeting on Phonology*. Washington, D.C.: Linguistic Society of America.

SLATER, ANNE S.; and SAUL FEINMAN. 1985. Gender and the phonology of North American first names. *Sex Roles* 13.429–440.

SMITH, JENNIFER L. 2016. Segmental noun/verb phonotactic differences are productive too. *Proceedings of the Linguistic Society of America*, ed. by Patrick Farrell, 1:article #3717. Washington, DC: Linguistic Society of America. http://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/3717.

STAMPE, DAVID. 1973. On chapter nine. *Issues in Phonological Theory*, ed. by Michael J Kenstowicz, Charles W Kisseberth, Michael J Kenstowicz, and Charles W Kisseberth, 44–52. The Hague: Mouton.

WILSON, COLIN; and MARIEKE OBDEYN. 2009. Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay.

WRIGHT, SAUNDRA K.; JENNIFER HAY; and TESSA BENT. 2005. Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics* 43.531–561.