Phonology **27** (2010) 205–224. © Cambridge University Press 2010
doi:10.1017/S0952675710000102

Underspecification and asymmetries in voicing perception[*]

So-One K. Hwang
University of Maryland

Philip J. Monahan
Basque Center on Cognition, Brain and Language

William J. Idsardi
University of Maryland

The purpose of our study is to show that phonological knowledge is an important basis for making predictions during speech perception. Taking the phonological constraint in English that coda obstruent clusters agree in their value for voicing, we conducted two experiments using vowel–stop–fricative sequences, where the task was to identify the fricative. Stimuli included sequences that were either congruent or incongruent. Consistent with models of featural underspecification for voiceless obstruents, our results indicate that only voiced stops induced predictions for an upcoming voiced fricative, eliciting processing difficulty when such predictions were not met. In contrast, voiceless stops appear to induce no equivalent predictions. These results demonstrate the important role of abstract phonological knowledge in online processing, and the asymmetries in our findings also suggest that only specified features are the basis for generating perceptual predictions about the upcoming speech signal.

1 Introduction

Phonological knowledge plays an important role in mapping surface phonetic forms to underlying representations (Halle 2002). One of the primary challenges for listeners is to undo the significant variation in the speech signal caused by speaker variation, coarticulation and phonological rule application, and ultimately arrive at the underlying linguistic representation of an utterance. Rarely, if ever, does a one-to-one relationship

[*] We would like to thank David Poeppel, Colin Phillips, Mathias Scharinger and the editors and reviewers of *Phonology* for valuable comments and advice. Thanks also to Brian Dillon, Diogo Almeida and Elika Bergelson for technical assistance. All remaining errors are the responsibility of the authors. This work was supported in part by grant NIH 7R01DC005660-07 to David Poeppel and William J. Idsardi.

exist between a surface acoustic cue and its corresponding phonological representation, as a given realisation of a phoneme can vary substantially along several phonetic continua (Liberman *et al.* 1967). Some models of speech perception have suggested that the analysis of the auditory input involves recovering the underlying representation and the set of generative phonological rules that applied in the form under analysis (Stevens & Halle 1967, Poeppel *et al.* 2008).

In this paper, we aim to show that phonological knowledge is an important basis for making predictions about the upcoming incoming speech signal during processing. The present study examines the phonological perception of voicing assimilation. Tautosyllabic English obstruent consonant clusters must agree in their specification of voicing (Greenberg 1978, Harms 1978, Mester & Itô 1989). This observation appears to be true of most languages that have obstruent consonant clusters and is true of all known coda clusters.[1] In English, this phonotactic generalisation holds in morphophonological alternations, as well as in monomorphemic obstruent consonant clusters. Using vowel–stop–fricative sequences, we investigate the influence of voicing features (i.e. [+voice] or [−voice]) on stop consonants and their effect on the perception of voicing on a following fricative. If auditory processing involves a mechanism for making predictions online, then we expect responses to segments that violate those predictions to show perceptual difficulty (manifested in lower accuracy and slower reaction times).

We compare two views on how voicing is represented featurally, each making different predictions for how obstruent clusters are actively processed during speech perception. If [+voice] and [−voice] are symmetrically encoded as binary features (as specified in rule 11 of Chomsky & Halle 1968: 178, for example), then both voiced and voiceless stops should affect the processing of a subsequent obstruent. Under this account, obstruent clusters that agree in voicing should be processed faster and more accurately than those that do not. Alternatively, if [voice] is a privative feature and [−voice] is underspecified (Mester & Itô 1989, Lombardi 1991, Avery & Idsardi 2001), only voiced obstruents should have predictive import in the processing of an immediately following obstruent. In contrast, voiceless stops are predictive of neither voiceless nor voiced fricatives. This article presents two experiments testing these hypotheses.

An important component of understanding speech perception is to identify which acoustic cues are informative in phonological processes. Theories of underspecification predict that certain features will have more import than others (Archangeli 1988, Steriade 1995), causing asymmetries in processing (Lahiri & Reetz 2002, Eulitz & Lahiri 2004, Obleser *et al.* 2004, Friedrich *et al.* 2006, Lahiri & Reetz 2010). Before describing

[1] Hebrew *onset* clusters appear to be a notable exception to this cross-linguistic generalisation. Examples include [kvarim] 'graves', [gfanim] 'vines', [tguva] 'reaction' (see Kreitman 2007).

our experiments, we first provide a brief background on theories of underspecification and describe how underspecification is thought to be employed in speech perception.

1.1 Underspecification

The most parsimonious phonological grammar is one where all and only the idiosyncratic properties are specified in the lexicon, and the predictable properties are derived via phonological rule application (Chomsky & Halle 1968). Theories of underspecification aim to accomplish this task by positing that all and only the marked or unpredictable features are stored for a given phonological segment, while the predictable feature values are supplied by phonological rules during the course of a derivation (Archangeli 1988; see Steriade 1995 for a discussion and thorough review of the various models of featural underspecification).

Perhaps the most commonly discussed underspecified feature is [coronal] (Avery & Rice 1989), particularly with respect to the nasal segment /n/. Typologically, the coronal nasal /n/ is far more likely to inherit the place of articulation of adjacent segments than its non-coronal nasal counterparts (e.g. /m/ or /ŋ/). Rarely do we find attested cases of non-coronal nasals assimilating to a coronal place of articulation. For example, in English, the coronal /n/ often assimilates to labial or dorsal: /θɪn bʊk/ → [θɪm bʊk] *thin book*. Cases of non-coronal nasals assimilating to coronal place of articulation: /brɪŋ tep/ → *[brɪn tep] *bring tape* are unattested, or at least extremely rare. Phonological segments that are underspecified for Place in their underlying representation are provided with a surface specification for place of articulation by phonological rule application.

In the literature, there have been two primary models of underspecification proposed for phonological feature inventories: contrastive underspecification and radical underspecification. Models of contrastive underspecification propose that only the features necessary to distinguish phonological segments in a language are specified, while features that do not serve a contrastive role between two segments are underspecified (Steriade 1987, Clements 1988). Radical underspecification, on the other hand, proposes that feature values that can be supplied via phonological rule application are not specified in a given segment's underlying representation (Archangeli 1988). For instance, in the coronal nasal assimilation example above, the place of articulation of the coronal nasal can be supplied via phonological rule application (spreading of place of articulation features), and therefore [coronal] does not need to be specified in the underlying featural representation for /n/. Although our study is not designed to tease apart these different approaches (Steriade 1995), their common claim about underspecification at the phonological level, as distinct from how features may be represented phonetically, is relevant to our discussion.

Lombardi (1991, 1995, 1999) re-examines previous accounts of the typology of laryngeal processes, and proposes to eliminate laryngeal distinctions within obstruent clusters. She claims that at least some phonological features are PRIVATIVE, specifically those for voicing and other laryngeal features. This model is to a first approximation equivalent to an underspecification account, where only [+voice] is specified and [−voice] is underspecified (in other words, [−voice] is not marked at the featural level and voicing is simply marked as [voice]). Although Lombardi (1991) makes a distinction between models of underspecification and privative features, we believe that they are compatible, given the assumption that sounds underspecified for [voice] do not have a laryngeal dimension in feature geometry. Dresher *et al.* (1995) and Avery & Idsardi (2001) argue for a phonological model of underspecification in which features that mediate phonetic and lexical levels of representation are either null (∅) or marked. This basis for contrast results in representational economy, where it is the dimensions of features that are contrastive rather than the gestures themselves. All of these theories share the common property that some phonetic features are not phonologically represented and use the lack of a featural value to provide theoretical accounts for the observed asymmetries in the typology of voicing phenomena.

1.2 Underspecification in processing

In the experiments reported here, we find asymmetries in the online processing of non-word stimuli. Although our results cannot be purely attributed to lexical effects, because they employ non-words and therefore do not directly bear upon theories about an underspecified lexicon, in this section we provide an overview of some important processing studies, based on underspecified lexical representations. In doing so, we show how underspecification theories can predict processing asymmetries.

Lahiri & Marslen-Wilson (1991) argue for lexical underspecification based on a gating study of nasal vowels in Bengali. In Bengali, nasal vowels have phonemic status. However, in monosyllabic words, only three of the four possible combinations of vowels and following consonants are phonetically attested: [CVC], [CṼN] and [CṼC]; no words contain *[CVN]. In contrast, in English, nasal vowels are predictable, occurring only before nasal consonants, giving only [CVC] and [CṼN]. Using a lexical gating task, Lahiri & Marslen-Wilson show that English speakers exploit their knowledge of nasality in vowels as predictive of a following nasal consonant, but Bengali speakers do not. They conclude that, because Bengali speakers do *not* treat [CṼ…] gates as ambiguous between [CṼN] and [CṼC], they must be lexically representing nasal vowels without the feature [nasal] before nasal consonants. The difference between the groups in the gating study is attributed to underspecified lexical representations: since [CṼN] words are lexically represented as /CVN/, they are poor matches to the nasality present in the [CṼ…] gates, whereas /CṼC/ items are good matches to the incoming input. For English speakers, however, the nasality cannot be attributed to the vowel in lexical representations;

therefore, the presence of nasality on vowels is always informative about the following nasal consonant.

Lahiri & Reetz (2002), building on studies such as Lahiri & Marslen-Wilson (1991), which find experimental evidence for lexical underspecification, differentiate among three possible outcomes in the mapping of features onto representations: MATCH, MISMATCH and NO-MISMATCH. A MATCH condition occurs if the signal and the underlying form have the same features, and a MISMATCH occurs if they have contradicting features. A NO-MISMATCH condition occurs if there is neither a match nor a contradiction of features, that is in cases where a feature is underspecified in the underlying form and feature matching cannot be completely evaluated. As an example, when the feature [labial] is present in the signal and is also in the underlying representation a MATCH occurs; when the feature [coronal] is present in the signal when the underlying representation is [labial] a MISMATCH occurs; however, when the feature [labial] is present in the signal when the underlying representation is underspecified for place a NO-MISMATCH occurs. Thus, in their model of a featurally underspecified lexicon, three different processing effects are predicted.

While specified features such as [labial] can induce place assimilation on underspecified segments, the reverse is unattested. Using this asymmetry in place assimilation, Lahiri & Reetz (2002) show that, behaviourally, non-coronal segments are tolerated as variants of underspecified coronal segments, but not *vice versa*. In a semantically primed lexical decision task, where German words like *Ho*[n]*ig* 'honey' primed *Biene* 'bee' and *Ha*[m]*er* 'hammer' primed *Nagel* 'nail', pseudo-word variants such as **Ho*[m]*ig* continued to prime *Biene*, but variants such as **Ha*[n]*er* did not continue to prime *Nagel*. Using electro-encephalography (EEG), Friedrich *et al.* (2006) showed that EEG components (i.e. N400), as well as behavioural measures, reflected differences in the activation of lexical items in a speeded lexical decision task. Pseudo-words such as **Pro*[d]*e*, a variant of *Pro*[b]*e* 'test', were more easily rejected as non-words than **Hor*[b]*e*, a variant of *Hor*[d]*e* 'horde'. Thus, although lexical representations of words containing the medial coronal consonants remained activated by a corresponding non-coronal consonant, lexical representations of words containing non-coronal consonants did not remain activated by a coronal consonant. This asymmetry suggests that coronal consonants are underspecified for place, whereas non-coronal consonants have specified features in the lexicon and are not vulnerable to assimilation in a similar manner. Friedrich *et al.* (2006) show that such knowledge is employed in online processing to make predictions during lexical access.

1.3 Phonological knowledge in online processing

In this section, we provide a summary of previous studies that have examined the role of phonological knowledge in speech perception.

Although they do not provide any discussion of underspecification, a close examination of their findings also reveals processing asymmetries compatible with such an account.

Previous results have suggested that the perception of English nasalised vowels triggers an explicit prediction for an upcoming nasal consonant. This prediction can be made in English, because nasal vowels, which are predictable allophonic variants of oral vowels, only occur preceding nasal consonants. In Fowler & Brown (2000), materials were created by taking natural disyllables such as [baCə] or [bãNə] and either splicing or cross-splicing them to create sequences of vowels and consonants that were congruent (VC, ṼN) or incongruent (ṼC, VN), with respect to nasalisation. This differs from Lahiri & Marslen-Wilson (1991), because participants received illicit sequences as well as licit ones. The task of the participants was to identify the consonant. The reaction times to the congruent VC and ṼN cases were faster than to the incongruent ṼC and VN cases. Moreover, they found a significant difference within the incongruent stimuli, with reaction times to ṼC being faster than to VN. The greatest processing difficulty was found for a sequence without proper anticipatory nasalisation on the vowel (VN) than for false nasalisation (ṼC).

Using magneto-encephalography (MEG), Flagg *et al.* (2006) tested VCV sequences that also agreed or differed in their specification for nasality (congruent: [aba], [ãma]; incongruent: [ãba], [ama]). Electrophysiological latencies (approximately 70 ms post-onset of the consonant) were overall shorter for the congruent than incongruent sounds, but the congruent VC sequence elicited faster latencies than other sounds, including the congruent ṼN sequence. The significant difference between the congruent pairs was one that was not found in Fowler & Brown (2000).

These studies show that although oral vowels are not specified for [oral] or [−nasal] *in the lexicon* for English, as proposed by Lahiri & Marslen-Wilson (1991), they are informative in *phonological processes*, and form the basis for predicting an upcoming oral consonant. These findings also suggest that English listeners are able to use the phonetic information contained in the vowel, together with their knowledge of the phonological sound patterns of English, to predict the feature of the upcoming consonant.

1.4 Voicing assimilation in online processing

In the present studies, we exploit the following cross-linguistically attested constraint: coda obstruent consonant clusters must agree in voicing. Stop–fricative clusters are commonly seen in the allomorphic variation of English plural formation. The plural marker is realised as [z] when following voiced stops, as in [dɔgz] *dogs*, and as [s] when following voiceless stops, as in [kæts] *cats*. The pattern also holds for the present 3rd person singular inflection and the possessive marker and auxiliary contraction. This phonotactic generalisation holds in morphophonemic alternations in English, as well as in monomorphemic obstruent consonant clusters (e.g. [læps] *lapse*). Given this pattern of voicing agreement within obstruent clusters, an attractive hypothesis would be that the congruent, empirically attested clusters should be processed faster and more easily than incongruent, empirically unattested clusters. Reaction times should be faster and accuracies higher for grammatical sequences, whereas reaction times should be slower and accuracies lower for ungrammatical sequences.

Theories of underspecification, however, predict asymmetric consequences for processing (Lahiri & Reetz 2002, Obleser *et al.* 2004, Friedrich *et al.* 2006). In the same way that [coronal] is not marked for place of articulation, as proposed in Lahiri & Reetz (2002) and demonstrated in Friedrich *et al.* (2006), the proposal for voicing is that [−voice] is not a feature represented in the phonological system (Lombardi 1991, Avery & Idsardi 2001). Combining Lombardi's representational proposal with the processing account of Friedrich *et al.* reviewed above, we predict asymmetric behaviour in the perception of voicing in obstruent clusters.

Here, we compare non-word syllables with coda obstruent clusters that agree in voicing (UDZ: [ubz udz ugz]; UTS: [ups uts uks]) with those that do not (UDS: [ubs uds ugs]; UTZ: [upz utz ukz]). When adopting an underspecification account for voicing (i.e. [voice] is privative, so that only voiced consonants are specified for [voice]) to online processing, we suspect that only voiced obstruents can be predictive of the following obstruent, whereas the lack of voicing on the stop consonant cannot induce predictions. Accordingly, we might expect to find two different behavioural effects. First, reaction times and accuracies should be facilitated when meeting a prediction for voicing from a voiced stop. Second, reaction times should increase and accuracies decrease when the prediction is violated by encountering a voiceless obstruent. On the other hand, neither facilitation nor difficulty should be observed when a voiceless stop is followed by a voiced or voiceless fricative, given the hypothesis that voiceless obstruents are underspecified for the feature [voice] and consequently cannot be exploited for the basis of online predictions. Thus, given this analysis, we predict the following scale of processing difficulty: UDZ is easier to process than UTS and UTZ, both of which are easier to process than UDS.

2 Experiment 1

In Experiment 1, we tested the alveolar stop–fricative clusters: [uts], [udz], [utz] and [uds]. The use of non-words in this design helps control for lexical biases that might favour congruent sequences of sounds and avoid other lexical factors, while still providing a way to investigate the role of phonological knowledge on the perception of a sequence of sounds. Speech perception with non-word stimuli also makes it possible to test to what degree features may be specified at lower levels of representation (phonological and phonetic), despite underspecification at the level of the lexicon. Although Lahiri & Marslen-Wilson (1991) show that lexically specified features play an important role in lexical tasks, surface phonetic realisations, such as allophonic variation, have also been shown to play an important role in phonological processing. As in our own experiments, Fowler & Brown (2000) used non-word stimuli in a consonant-identification task, in which English speakers were asked to identify if the consonant was oral or nasal after hearing an oral or nasal vowel (unlike the studies of Lahiri & Marslen-Wilson 1991 and Friedrich *et al.* 2006, which used lexical gating or lexical decision tasks to investigate processing).

If voicing is encoded as a binary feature ([±voice], the symmetric hypothesis), then incongruent clusters ([tz], [ds]) should result in slower reaction times and lower accuracy than congruent clusters ([ts], [dz]). In contrast, if voicing is privative ([voice], the asymmetric hypothesis) and if only specified features can serve as the basis for phonological predictions, then [dz] should elicit the fastest reaction times and highest accuracy, and [ds] should elicit the slowest reaction times and lowest accuracy.

Participants. Ten native speakers of American English (six female; age range 18–21), who were naive as to the purpose of the experiment, participated in Experiment 1. Based on the results of a subject outlier analysis, two participants were excluded from further analysis, due to exceedingly short reaction times (< 600 ms post-stimulus onset). Because reaction times are calculated from the stimulus onset and the relevant information from the fricative is not available until 200 ms into the signal, these exceedingly short reaction times suggest that the participants had been initiating motor responses before they heard the relevant part of the stimuli. All participants provided written informed consent and either received course credit or were paid for their participation.

Stimuli. A male native speaker of American English recorded natural VC_1C_2 utterances of [uts] and [udz]. These recordings were edited using Praat (Boersma & Weenink 2008), so that each phonetic segment was 100 ms in duration and the total duration of all stimuli was 300 ms. When the naturally recorded segments were shorter than 100 ms, as was the case with stops, they were made longer by copying periodic sections at zero-crossings in medial portions of the segments and pasting them next to the copied sections; any required partial pitch period was placed immediately before the burst. When the naturally recorded segments were longer than 100 ms, as was the case with the vowels and fricatives, they were made shorter by excising the beginning of the vowel and the end of the fricatives, because these portions were going to be ramped in later processing. This procedure ensured that the edits had no effect on the segment transitions. The resulting 300 ms items were gradually ramped so that V had a 20 ms fade-in and C_2 had a 20 ms fade-out. The final stimulus tokens

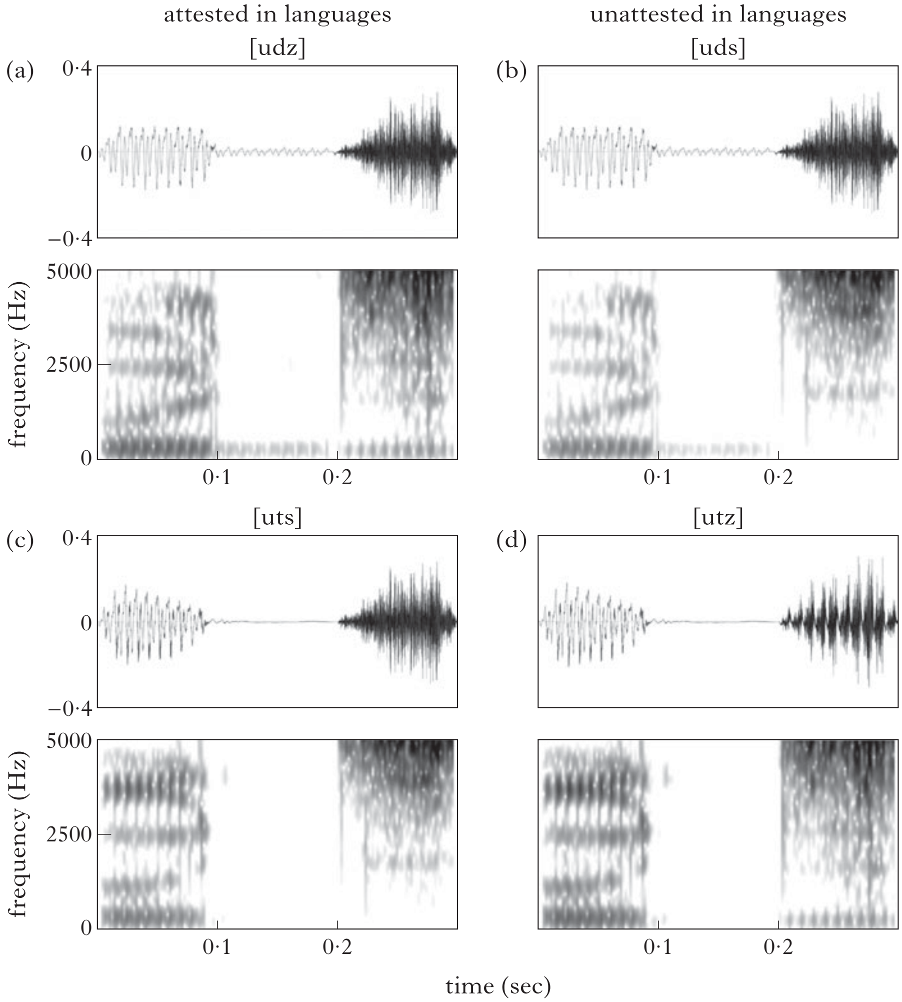


Figure 1

Waveforms and spectrograms of stimuli used in Experiment 1: (a) [udz], (b) [uds], (c) [uts], (d) [utz]. The amount of voicing is evident by the periodicity in the waveforms and the low-frequency energy in the spectrograms between 0·1 and 0·2 sec in (a) and (b) and their respective absence in (c) and (d).

were created by cross splicing a final fricative ([s] or [z]) with a vowel–stop sequence ([ut] or [ud]) to create tokens with voicing agreement (i.e. [uts] and [udz], the congruent cases) and voicing disagreement (i.e. [utz] and [uds], the incongruent cases). Copies of these four items were made to create versions without the stop-release bursts. The stop-release bursts (at the end of C_1) were replaced by silence, maintaining the 100 ms duration

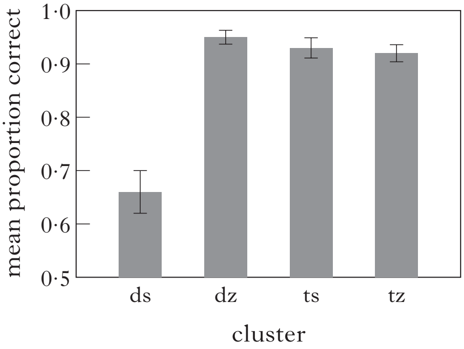


Figure 2
Mean proportion of correct responses for each cluster in Experiment 1. Error bars indicate one standard error of the mean.

of the segment.[2] The editing of the material in this way was done so that the same material could be used in a companion MEG experiment, which required precise time-locking in the stimuli material (Hwang *et al.* 2009). Thus there were eight items in this study, and stimulus presentation included 150 randomised trials of each of the eight tokens, using Presentation (Neurobehavioral Systems, Inc.) software. No filler items were used, resulting in a total of 1200 items. The waveforms and spectrograms used are given in Fig. 1.

Procedure. The experiment was conducted in a single testing session separated into three blocks. The participants sat facing a computer screen in a sound-attenuated room, wearing a headset. The stimuli were presented at equal volumes to both ears at a comfortable intensity level. The participants were instructed to keep their hands on the keyboard and to use their index fingers to press the 'F' or 'J' keys, depending on whether they heard [z] or [s]. Response and key arrangements were counterbalanced across participants. The participants were instructed to respond as quickly and accurately as possible. The 150 repetitions were divided into three blocks of 50 items each. The interstimulus interval was randomised between 1250 ms and 1750 ms. After each block, the participants were given a brief self-timed break; they pressed the space bar to continue on to the next block. No feedback was given regarding the responses during the sessions, although the nature of the experiment was discussed with the participants during the debriefing period after its conclusion. Each session lasted approximately 40 minutes.

Accuracy results. Fig. 2 shows the mean proportion of correct responses for each cluster. The data from all experiments were analyzed

[2] The stimulus materials are available (April 2010) at http://www.ling.umd.edu/~idsardi/materials/.

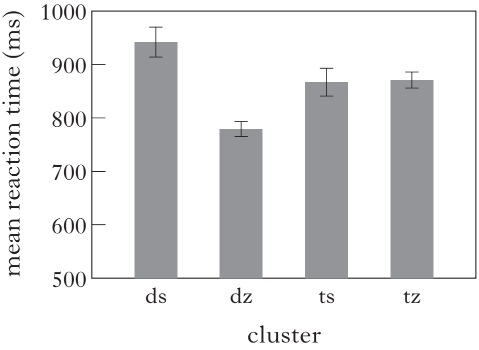


Figure 3

Mean reaction times for correct responses for each cluster in Experiment 1. Error bars indicate one standard error of the mean.

using the JMP 6 and 7 statistical packages (SAS Inc.) and R 2.8.1 (R Development Core Team 2005).

General linear mixed effects models (logistic regression with binomial errors, with Subject as a random effect) were calculated on the correct and incorrect counts for nested models of fixed effects using the nlme package for non-linear mixed effects in R (Crawley 2007, Baayen 2008, Pinheiro *et al.* 2009). Model comparison showed that the simple model adequately covered the data with only Cluster as a main effect; no significantly greater coverage was achieved by including terms for burst presence or for button arrangement. This was confirmed by examination of the analysis of variance table (ANOVA) for the full model in which Cluster was significant ($F(3,48)=50{\cdot}36$, $p<0{\cdot}0001$) and all other main effects and interactions were non-significant (all $F<1$). Tukey-Kramer Honestly Significant Differences (HSD) post hoc multiple comparison tests ($\alpha=0{\cdot}05$) for the clusters showed that responses to [ds] were significantly less accurate than those for the other three clusters. However, although [dz] had the highest mean accuracy, it was not significantly more accurate than [ts] or [tz], most likely due to a ceiling effect in the accuracy of responses to these clusters.

The asymmetric hypothesis correctly predicts the rank order of the accuracies and the fact that [ds] is significantly less accurate than the other conditions. Moreover, the trend for [dz] is in the right direction, but does not reach statistical significance. These results are not consistent with the symmetric hypothesis, because the response to the ungrammatical [ds] cluster is significantly different from the response to the other ungrammatical [tz] cluster.

Reaction time results. Figure 3 shows the mean reactions times in milliseconds for correct responses for each cluster type. General linear mixed effects models (ANOVA, with Subject as a random effect) were calculated on the log-transformed reaction time data for correct responses for nested models of fixed effects. Model comparison showed again that the simple model provided adequate coverage of the data with only Cluster as a main effect; no significantly greater coverage was achieved by including terms for burst presence or for button arrangement. This was confirmed by examination of the ANOVA table for the full model, in which Cluster was significant ($F(3,48)=17{\cdot}91$, $p<0{\cdot}0001$) and all other main effects and interactions were non-significant (all $F<1$). Tukey-Kramer HSD post hoc tests ($\alpha=0{\cdot}05$) showed that responses to [ds] were significantly slower and that responses to [dz] significantly faster than the others, but that responses to [tz] and [ts] were statistically indistinguishable. These results fully support a processing interpretation of the asymmetric hypothesis that only [+voice] is marked and that [−voice] is featurally underspecified, and therefore only voiced segments induce predictions about following consonants.

3 Experiment 2

In Experiment 2, we sought to replicate the results from Experiment 1 and to extend them to the other contrastive places of articulation for stops. Experiment 2 tests the labial and velar places of articulation in addition to replicating the alveolar series. Based on the findings of Experiment 1, we chose to include only the stimuli where the stop-release bursts had been removed, again to gain a set of stimuli that could be used in a companion MEG experiment. Overall, we predict that the results will support the asymmetric hypothesis: UDS ([ubs uds ugs]) should elicit the most difficulty (lowest accuracy, highest reaction time), while UDZ ([ubz udz ugz]) should elicit the least (highest accuracy, lowest reaction time). Moreover, if voiceless obstruents are underspecified for voicing, then we predict no difference between UTS ([ups uts uks]) and UTZ ([upz utz ukz]) stimuli with a voiceless stop in C_1 position.

Participants. Twelve native speakers (ten female; age range 18–21) of American English, who were naive as to the purpose of the experiment and did not take part in Experiment 1, participated in Experiment 2. Based on the results of a subject outlier analysis, two participants were excluded from further analysis, due to exceedingly short reaction times (<600 ms), for the same reasons as described in Experiment 1. Thus we report the data from ten participants. All participants provided written informed consent and either received course credit or were paid for their participation.

Stimuli. In addition to the recordings used in Experiment 1, the same male native speaker of English recorded natural utterances of [ups], [ubz], [uks] and [ugz]. All sounds were edited in the same manner as in Experiment 1 to create new congruent and incongruent cluster tokens, and the stop-release bursts were removed. Participants heard 150 randomised trials of each of the twelve sounds, for a total of 1800 items.

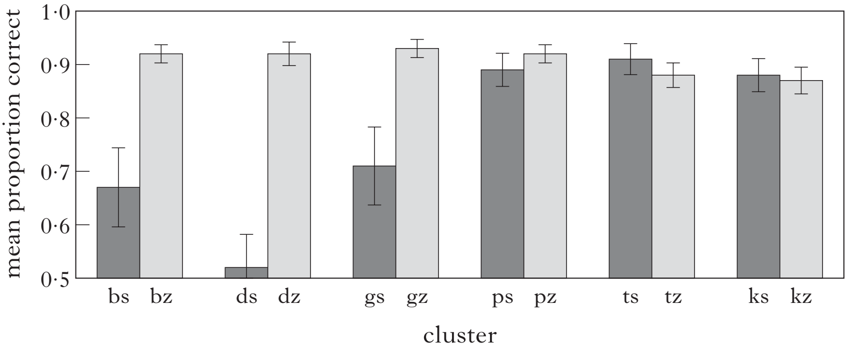


Figure 4

Mean proportion of correct responses for each cluster in Experiment 2. Error bars indicate one standard error of the mean.

Procedure. The procedure was identical to Experiment 1. Each session lasted approximately one hour.

Accuracy results. Based on the results from Experiment 1, we predicted that DS clusters would be the most difficult and that DZ would be the easiest. Figure 4 shows the mean proportion of correct responses for each cluster for the remaining ten participants. General linear mixed effects models (logistic regression with binomial errors, with Subject as a random effect) were calculated on the correct and incorrect counts for nested models of fixed effects. Model comparison and examination of the ANOVA tables showed a significant main effect for cluster voicing ($F(3,99)=51{\cdot}1361$, $p<0{\cdot}0001$), and a marginally significant interaction with the place of articulation of the first consonant ($F(6,99)=2{\cdot}5276$, $p<0{\cdot}03$), but no main effect for place of articulation ($F(2,99)=2{\cdot}2431$, $p>0{\cdot}11$), and no main effects or interactions for button arrangement (all $F<1$). Tukey-Kramer post hoc tests ($\alpha=0{\cdot}05$) showed that responses to UDS clusters were significantly less accurate than the others; the interaction with place of articulation was due to [ds] being significantly worse than [gs] ($p<0{\cdot}01$), and marginally worse than [bs] ($p<0{\cdot}1$), perhaps suggesting some additional difficulty in processing homorganic clusters. In summary, the accuracy results in Experiment 2 replicate the findings of Experiment 1: DS clusters are less accurately perceived than other clusters.

Overall, these results fully replicate our findings in Experiment 1. We found that DS clusters were significantly less accurate than all other cases, including TZ. As already noted, this finding is consistent with the asymmetric hypothesis based on an underspecification account, which claims that only voiced obstruents induce predictions regarding the voicing of upcoming obstruents. In contrast, the symmetric hypothesis predicts that DS and TZ should pattern similarly, a prediction not borne out in our results.

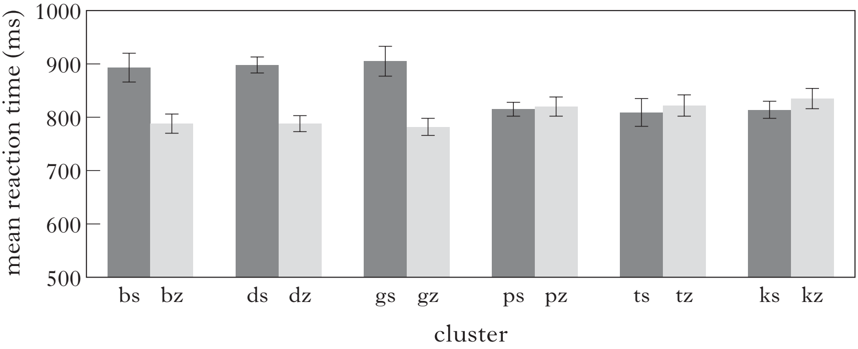


Figure 5

Mean reaction times for correct responses for each cluster in Experiment 2. Error bars indicate one standard error of the mean.

Reaction time results. Figure 5 shows mean reaction times (in ms) for correct responses for each cluster in Experiment 2. General linear mixed effects models (ANOVA, with Subject as a random effect) were calculated on the log-transformed reaction time data for correct responses for nested models of fixed effects. Model comparison and examination of the ANOVA tables showed a significant main effect for cluster voicing (F(3,103) = 37·1549, p < 0·0001), but no significant main effects for place of articulation or button arrangement (both F < 1). Interactions effects were also non-significant (voicing by button arrangement barely approached significance : F(3,6) = 3·2127, p = 0·1041 ; all other F < 1). Planned comparisons based on the asymmetric hypothesis showed that responses to DS were slower than those to TZ and TS (F(1,107) = 61·94, p < 0·0001), and responses to DZ were faster than those to TZ and TS (F(1,107) = 11·97, p < 0·001). Thus both aspects of the asymmetric hypothesis were confirmed in the reaction time measures. Figure 6 shows the reaction times for Experiment 2 grouped by cluster type.

In summary, Experiment 2 replicated the results from Experiment 1 by showing that reaction times to DS were much slower and that accuracy rates were much lower than for any other clusters. The data also showed that reaction times to DZ were much faster and accuracy rates higher than for the other clusters. Based on these findings, we can conclude that results from Experiment 1 can be extended to all places of articulation. Again, the findings of asymmetric processing are consistent with processing interpretations of voicing underspecification.

4 Discussion

Our findings demonstrate that a symmetric analysis, whereby both [+voice] and [−voice] are represented and thereby both serve as the basis

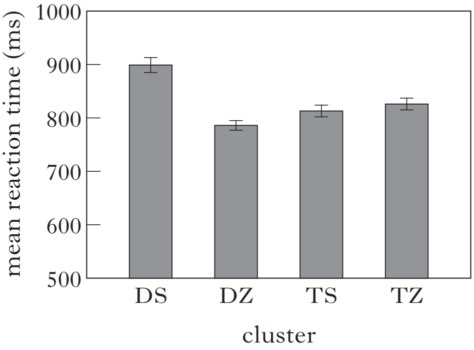


Figure 6

Mean reaction times for correct responses for each cluster type in Experiment 2. Error bars indicate one standard error of the mean.

for the generation of predictions in online speech perception, does not fully account for the accuracy and timing of consonant identification in clusters. In particular, we found that participants recognise UDS clusters more slowly and less accurately than other clusters, particularly UTZ. Moreover, they recognise UDZ more quickly than the other clusters, including UTS. The fact that we found differences within the congruent clusters (UDZ *vs.* UTS), as well as within the incongruent clusters (UDS *vs.* UTZ), suggests that a general, unified constraint against voicing disagreement in obstruent clusters, such as AgreeObs[Voice] (van Rooy & Wissing 2001), is not adequate to explain our results. These asymmetric results are more consistent with underspecification theories, according to which only specified features have import in phonological processes (Lombardi 1991, 1995, 1999, Avery 1996, Avery & Idsardi 2001, etc.).

These results also suggest that listeners did not treat UTS and UTZ sequences differently, although the former obeyed and the latter violated the constraint on voicing assimilation. We might attempt to account for these findings based on phonotactically triggered misperception, where non-native sequences are misperceived as native ones (Massaro & Cohen 1983, Berent *et al.* 2008). However, the hypothesis that ungrammatical sequences are heard as grammatical ones is not compatible with our result, which showed that listeners treat UDZ and UDS sequences differently. Moreover, we found that listeners do not treat all illegal clusters in the same way. Although both UTZ and UDS involve voicing disagreement, only the latter caused processing difficulty. Voiced fricatives, when perceived in the context of no predictions (following voiceless stops), cause neither facilitation nor difficulty. Thus, neither a symmetric analysis based on equal specification of [+voice] and [−voice] nor a phonotactic account can explain our asymmetric results. In short, voiced stops induce a prediction about the voicing of the following fricative, facilitating the recognition of a following voiced obstruent but interfering with the recognition of a following voiceless obstruent.

Underspecified representations have been hypothesised to occur at different levels of representation: lexical, phonological and phonetic (Keating 1988). With respect to these three levels, there are four logical possibilities that one could consider (see Table I). Because underlying representations must be recovered from surface information, we can make a directional hypothesis that underspecification at lower levels entails underspecification at higher levels, but not *vice versa*.

	A	B	C	D
lexical	✓	∅	∅	∅
phonological	✓	✓	∅	∅
phonetic	✓	✓	✓	∅

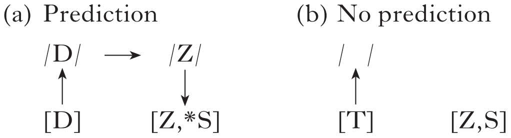
Table I

Possibilities for underspecification in voicing among the levels of representation. ✓ means that [−voice] is specified; ∅ that[−voice] is underspecified.

Based on the asymmetry in our results, we can reject possibility A, where [−voice] is specified at each level of representation. Because we used non-word sound sequences in our study, our results cannot be purely attributed to underspecification in the lexicon (possibility B). Thus, we only seriously consider possibilities C and D to explain our results. Both hypotheses support a theory of phonological underspecification. In this discussion, we suggest that our results support theories of underspecification at the phonological level, while support for phonetic underspecification is inconclusive.

We believe that online predictions about upcoming segments are mediated through phonology and do not occur at the phonetic level alone. We consider the possibility that top-down effects in processing may be driven by the listener's experience with the statistical distribution of the sounds that involves an analysis of the sounds at only the phonetic level. However, this frequency-based account does not extend to our results. In fast, natural speech, /z/ is often devoiced in American English, yielding final clusters, such as [ds] (Ohala 1983, Smith 1997). However, the incongruent combination of [tz] is rarely, if ever, attested. Nevertheless, the UDS cluster had the lowest accuracy and slowest reaction times. Thus, in the same way that models based on phonotactically triggered misperception cannot adequately distinguish the asymmetric treatment of UDS and UTZ sequences, attributing listener's expectations about upcoming segments solely to surface phenomena faces many challenges.

We provide the proposal in Fig. 7 for how online predictions about upcoming segments is mediated through a phonological level at which [−voice] is underspecified.



(a) Prediction

(b) No prediction

Figure 7

Schematic of a model for how online predictions about upcoming segments is mediated through phonology. In (a), sounds specified for [+voice] are informative in phonological processes and lead to predictions about the upcoming segment, whereas in (b), sounds with the unspecified feature [−voice] are not predictive. * indicates the violation of a prediction.

In Fig. 7, a voiced stop is directly mapped to an underlying representation where voiced sounds are specified. A sound that is specified for [voice] induces a strong phonological prediction that the following fricative is also marked for [voice], thus making a phonetic prediction of an upcoming voiced segment like [z]. Our results show facilitation, as indexed by faster reaction times, when phonological predictions are met. However, when this prediction for an upcoming voiced segment fails, we find evidence of processing difficulty, as indexed by lower accuracies and longer reaction times in recognising the voiceless fricative. The lack of voicing on a voiceless stop does not directly map to an underlying representation for a voicing feature; rather this information becomes underspecified. In our perceptual model, underspecified features are not informative in phonological processes and do not induce predictions about an upcoming segment. Thus, voiced and voiceless fricative are treated equally following underspecified stops. Our results show that underspecified stops do not induce predictions for a following underspecified fricative. Without a strong phonological prediction about the upcoming segment, voiced [z] and voiceless [s] are treated in the same way at both the phonological and phonetic levels.

Although Lahiri & Reetz's (2002) ternary choice of MATCH, MISMATCH and NO-MISMATCH conditions applies to a model for an underspecified lexicon, it can be extended to describe the results shown in our experiments. In condition (a) in Fig. 7, a MATCH condition is achieved when a voiced stop is predictive of a following voiced fricative and is followed by a voiced fricative, and a MISMATCH occurs when it is followed by a voiceless fricative, contrary to the prediction induced by the voiced stop. Condition (b) demonstrates the case of a NO-MISMATCH, where a voiceless stop does not induce phonetic predictions about an upcoming segment, and subsequent voiced and voiceless fricatives are treated equally. The asymmetries in our findings and the three-way division in the results are compatible with this type of analysis.

It is possible to envisage other accounts of our findings that would maintain a representational system where [+voice] and [−voice] are treated equally. Asymmetries present in lower-level auditory processes could potentially manifest themselves in reaction times and accuracy measures in consonant identification. In particular, if signal periodicity detection is asymmetric, such that periodic signals facilitate the identification of following periodic signals, while aperiodic signals do not, this could underlie the voicing decision that we find in the present data as periodicity is a significant correlate of obstruent voicing. However, other similar experiments on nasality (Fowler & Brown 2000, Flagg *et al.* 2006) also report asymmetric results and would require a different auditory explanation. Specifically, Fowler & Brown (2000) found an asymmetric pattern of responses in the ungrammatical sequences, such that a nasalised vowel followed by an oral consonant elicited faster reaction times than an oral vowel followed by a nasal consonant. Additionally, Flagg *et al.* (2006), in an electrophysiological experiment, found an asymmetrical pattern in the neural responses within the grammatical sequences: oral vowels followed by oral consonants elicited faster neuromagnetic responses than nasalised vowels followed by nasal consonants. In the absence of specific linking hypotheses for attested auditory processes that could underlie these asymmetries, it is more reasonable to believe that the experimentally observed asymmetries arise from within phonological representations.

5 Conclusion

Taking Lombardi's (1991) broader account that [voice] is a privative feature, in the present study we have provided evidence from the processing of non-word stimuli to show that voiceless obstruents are phonologically underspecified. These results provide an interesting point of comparison to the findings from nasal assimilation in Fowler & Brown (2000), where both the oral ([−nasal]) and nasal ([+nasal]) features on vowels are informative in predicting the upcoming consonant in non-word stimuli. Thus, despite the theory of lexical underspecification regarding [+oral]/[+nasal] features on vowels from Lahiri & Marslen Wilson (1991), it seems that such features may be specified at lower levels. The present findings are more consistent with asymmetric processing that was found for place assimilation in Friedrich *et al.* (2006) and Lahiri & Reetz (2002). Future work that connects grammatical models with processing studies will contribute to a better understanding of representational levels and the differences between them. Further investigations into the processing of speech sequences may reveal other asymmetries that have consequences for our understanding of how our phonological knowledge is represented. Moreover, online processing studies can further inform models of how features are phonologically encoded. These results demonstrate the important role of abstract phonological knowledge in online processing, and the asymmetries in our findings also suggest that only specified features are the basis for generating perceptual predictions about the upcoming speech signal.

REFERENCES

Archangeli, Diana (1988). Aspects of underspecification theory. *Phonology* **5**. 183–207.

Avery, Peter (1996). *The representation of voicing contrasts*. PhD dissertation, University of Toronto.

Avery, Peter & William J. Idsardi (2001). Laryngeal dimensions, completion and enhancement. In Hall (2001). 41–70.

Avery, Peter & Keren Rice (1989). Segment structure and coronal underspecification. *Phonology* **6**. 179–200.

Baayen, R. H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Berent, Iris, Tracy Lennertz, Jongho Jun, Miguel A. Moreno & Paul Smolensky (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences* **105**. 5321–5325.

Boersma, Paul & David Weenink (2008). *Praat: doing phonetics by computer* (version 5.0.08). http://www.praat.org/.

Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.

Clements, G. N. (1988). Toward a substantive theory of feature specification. *NELS* **18**. 79–93.

Crawley, Michael J. (2007). *The R book*. Chichester: Wiley.

Dresher, B. Elan, Glyne Piggott & Keren Rice (1995). Contrast in phonology: overview. *Toronto Working Papers in Linguistics* **13**. iii–xvii.

Eulitz, Carsten & Aditi Lahiri (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of Cognitive Neuroscience* **16**. 577–583.

Flagg, Elissa J., Janis E. Oram Cardy & Timothy P. L. Roberts (2006). MEG detects neural consequences of anomalous nasalization in vowel–consonant pairs. *Neuroscience Letters* **397**. 263–268.

Fowler, Carol A. & Julie M. Brown (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception and Psychophysics* **62**. 21–32.

Friedrich, Claudia K., Carsten Eulitz & Aditi Lahiri (2006). Not every pseudoword disrupts word recognition: an ERP study. *Behavioral and Brain Functions* **2:36**.

Greenberg, Joseph H. (1978). Some generalisations concerning initial and final consonant clusters. In Joseph H. Greenberg (ed.) *Universals of human languages*. Vol. 2: *Phonology*. Stanford: Stanford University Press. 243–279.

Hall, T. Alan (ed.) (2001). *Distinctive feature theory*. Berlin & New York: Mouton de Gruyter.

Halle, Morris (2002). *From memory to speech and back: papers on phonetics and phonology 1954–2002*. Berlin & New York: Mouton de Gruyter.

Halle, Morris & Kenneth N. Stevens (1962). Speech recognition: a model and a program for research. *IRE Transactions on Information Theory* **8**. 155–159.

Harms, Robert T. (1978). Some nonrules of English. In Mohammad Ali Jazayery, Edgar C. Polomé & Werner Winter (eds.) *Linguistic and literary studies in honor of Archibald A. Hill*. Vol. 2: *Descriptive linguistics*. The Hague: Mouton. 39–51.

Hwang, So-One K., Philip J. Monahan & William J. Idsardi (2009). Asymmetric phonological predictions in speech perception: MEG evidence. Paper presented at the 16th Annual Meeting of the Cognitive Neuroscience Society, San Francisco.

Keating, Patricia A. (1988). Underspecification in phonetics. *Phonology* **5**. 275–292.

Kreitman, Rina (2007). *The phonetics and phonology of onset clusters : the case of Modern Hebrew*. PhD dissertation, Cornell University.

Lahiri, Aditi & William Marslen-Wilson (1991). The mental representation of lexical form: a phonological approach to the recognition lexicon. *Cognition* **38**. 245–294.

Lahiri, Aditi & Henning Reetz (2002). Underspecified recognition. In Carlos Gussenhoven & Natasha Warner (eds.) *Laboratory Phonology 7*. Berlin & New York: Mouton de Gruyter. 637–675.

Lahiri, Aditi & Henning Reetz (2010). Distinctive features: phonological underspecification in representation and processing. *JPh* **38**. 44–59.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler & M. Studdert-Kennedy (1967). Perception of the speech code. *Psychological Review* **74**. 431–461.

Lombardi, Linda (1991). *Laryngeal features and laryngeal neutralization*. PhD dissertation, University of Massachusetts, Amherst.

Lombardi, Linda (1995). Dahl's law and privative [voice]. *LI* **26**. 365–372.

Lombardi, Linda (1999). Positional faithfulness and voicing assimilation in Optimality Theory. *NLLT* **17**. 267–302.

Massaro, Dominic W. & Michael M. Cohen (1983). Phonological context in speech perception. *Perception and Psychophysics* **34**. 338–348.

Mester, Armin & Junko Itô (1989). Feature predictability and underspecification: palatal prosody in Japanese mimetics. *Lg* **65**. 258–293.

Obleser, Jonas, Aditi Lahiri & Carsten Eulitz (2004). Magnetic brain response mirrors extraction of phonological features from spoken vowels. *Journal of Cognitive Neuroscience* **16**. 31–39.

Ohala, John J. (1983). The origin of sound patterns in vocal tract constraints. In Peter F. MacNeilage (ed.) *The production of speech*. New York: Springer. 189–216.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar & the R Core Team. (2009). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-93.

Poeppel, David, William J. Idsardi & Virginie van Wassenhove (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B* **363**. 1071–1086.

R Development Core Team (2005). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at http://www.r-project.org.

Rooy, Bertus van & Daan Wissing (2001). Distinctive [voice] implies regressive voicing assimilation. In Hall (2001). 295–334.

Smith, Caroline L. (1997). The devoicing of /z/ in American English: effects of local and prosodic context. *JPh* **25**. 471–500.

Steriade, Donca (1987). Redundant values. *CLS* **23:2**. 339–362.

Steriade, Donca (1995). Underspecification and markedness. In John A. Goldsmith (ed.) *The handbook of phonological theory*. Cambridge, Mass. & Oxford: Blackwell. 114–174.

Stevens, Kenneth N. & Morris Halle (1967). Remarks on analysis by synthesis and distinctive features. In Weiant Wathen-Dunn (ed.) *Models for the perception of speech and visual form*. Cambridge, Mass: MIT Press. 88–102.