English Language and Linguistics 17.2: 241–269. © Cambridge University Press 2013
doi:10.1017/S1360674313000026

Using nonsense words to investigate vowel merger[1]

JENNIFER HAY
New Zealand Institute of Language, Brain and Behaviour, University of Canterbury

KATIE DRAGER
University of Hawai'i at Mānoa

and

BRYNMOR THOMAS
United Arab Emirates University

(Received 28 June 2012; revised 22 November 2012)

In previous work, we have found that New Zealand listeners who produce merged tokens of NEAR and SQUARE can accurately distinguish between the vowels in perception even though they report that they are guessing. The ability to distinguish the vowels is affected by a variety of factors for these listeners, including the likelihood that the speaker and experimenter maintain the distinction (Hay *et al.* 2006b; Hay *et al.* 2010). In this article, we report on experiments that examine the production and perception of real and nonsense words in the context of two mergers: the ELLEN/ALLAN merger in New Zealand English and the LOT/THOUGHT merger found in American English. The results demonstrate that speakers' degree of merger depends at least partially on whether the word is a real or nonsense word. Additionally, the results indicate that a token's real word status affects the merger differently in production and perception. We argue that these results provide evidence in favour of a hybrid model of speech production and perception, one with both abstract phoneme-level representations and acoustically detailed episodic representations.

1 Introduction

When linguists discuss the status of a merger in an individual or a population, they tend to focus on the degree of acoustic overlap involved in producing the forms. It's quite possible, however, that even when forms are produced in an acoustically similar manner, an individual may classify them as belonging to different categories. Some listeners who produce merged word forms might not perceive any difference between minimal pairs, whereas others may easily distinguish between them in perception. Additionally, exposure is likely to play a role, so that degree of merger may vary between previously encountered words and words that have not been previously encountered. Thus, when we describe an individual as being 'merged' on the basis of their production, this could actually describe several different states. In this article, we look at the production and perception of real (previously encountered) and nonsense (novel) words in two examples of near-complete mergers (one conditioned, one unconditioned) in an attempt

[1] This work was supported by a Rutherford Discovery Fellowship to the first author. We are grateful to our referees for their detailed feedback, to Rebecca Clifford and Liam Walsh for their help with analysis, and to the New Zealand Institute of Language, Brain and Behaviour for their financial support.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

to understand the relationship between production and perception, and to explore the possible roles of episodic memories and their associated abstract categories.

2 Background

Most sociolinguistic work on mergers has focused on the mechanisms through which mergers take place: which speakers or groups of speakers lead or resist a merger-in-progress (Irons 2007), how mergers spread through a community and outward to new communities (Bailey *et al*. 1993; Boberg 2000; Irons 2007), and how sounds that are merging are linked with other sounds in a speaker's phonological system (Labov 2001, 2010). Labov *et al.* (1991) identify a number of near-mergers, in which speakers do not report hearing a difference despite a consistent difference in their production. In contrast, Hay *et al.* (2006b) found that individuals were highly accurate at identifying sounds undergoing merger, even though these same individuals merge the sounds when speaking and also feel they are guessing during the identification task. These results are intriguing and leave us asking: why are results across the different studies so different? And how is it possible for individuals to differ so much across their production and perception?

There is evidence that speakers and hearers access phonetically detailed episodic memories of words, with rich social indexing (Strand & Johnson 1996; Goldinger 1996; Hay *et al.* 2006b). There is also evidence that, in addition to episodic memories, more abstract representations exist (Nielsen 2011; Hay *et al.* 2010), i.e. categories that correspond to phonemes and that are based on generalizations over the episodic memories. If this is the case, then people who fully merge two sounds should have one category, and speakers who are fully distinct should have two. Furthermore, an individual in a pre-final state of the merger may also have two categories. Hay *et al.* (2010) outline a range of tasks in which individuals who merge two sounds to some degree were exposed to dialects in which the distinction between the sounds is maintained. The effect of this exposure varied depending on the degree of merger found in the speech of the individual and the type of task being completed; when compared with behaviour after exposure to merged varieties, merged listeners exposed to distinct varieties were less accurate in perception but produced more of a distinction in production, whereas less merged individuals' perception was not affected by the distinct varieties (though they also increased the amount of distinction made in production). Hay *et al.* (2010) argue that some individuals have a single phoneme representation for the sounds, whereas others have two, and that this causes different behaviours between the two groups. Furthermore, they argue that different tasks access different levels of representation, so that the difference between the groups of speakers should be observed in tasks, such as identification tasks, that access the phoneme level in addition to the phonetic representations.

Hay *et al.* (2010) focus on the degree to which different tasks access the phoneme level, but there is another way to manipulate the degree to which the phoneme level is accessed: by using words which do not have any episodic memories associated

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

with them. If both episodic memories and abstract generalizations are involved in production and perception, then we might predict different behaviours when episodic memories are absent and the abstract generalization is all that can be relied upon. In this article we explore the idea of manipulating lexical status. Do real words and non-words differ in their influence on vowels undergoing merger? We investigate this question with two nearly complete mergers, each found in different dialect areas. The first merger we consider is a conditioned merger which is near complete in New Zealand English: a merger of DRESS and TRAP before /l/ referred to as the ELLEN/ALLAN merger. Because this merger is conditioned, there is good reason to believe that individuals might associate the prelateral vowels with two categories for some time after they stop producing a distinction themselves. The two categories exist independently for the non-prelateral productions, and speakers may be able to access these distinct categories even though their phonetic representations are overlapping in pre-/l/ context. The second merger we consider is the merger of the back vowels LOT and THOUGHT in Hawai'i and American English. This merger is related to phonological context in some varieties of American English, but in Hawai'i it appears to be nearly complete in all phonological contexts. This means that speakers may be less likely to have two distinct categories, though some degree of distinction may nonetheless exist in their mental representations of sounds due to interactions with people who maintain a distinction.

For these two populations, we conducted simple production and perception experiments. The goal of the experimental tasks is to manipulate lexical status to assess whether our participants associate the vowels with one category or two, and to determine whether this association differs across production and perception.

2.1 The theoretical perspective

In exemplar-based models of speech production and perception, the mental representations of sounds are episodic, phonetically detailed, and constantly updated based on experience (Johnson 1997; Pierrehumbert 2001). Additionally, there is a direct link between representations of phonetic detail and representations of words. These models are motivated by evidence that listeners store considerably more phonetic detail than previously thought, producing very slight but consistent differences in production and demonstrating sensitivity to phonetic detail during perception (Goldinger 1996; Foulkes *et al.* 2010). Additionally, there is increasing evidence that at least some of this phonetic detail is stored by listeners in such a way that it is linked to representations of the word in which it was encountered (Goldinger 1996; Jurafsky *et al.* 2002; Aylett & Turk 2004; Hay & Bresnan 2006; Baker & Bradlow 2009). There is evidence that homophones and polysemes, which by definition share a phoneme-level representation, can in fact have systematically different phonetic realizations in production (Hay & Bresnan 2006; Gahl 2008; Drager 2011a). Furthermore, listeners can use this phonetic detail when identifying the words in perception (Drager 2010). Thus, sounds appear to be stored with phonetic detail and these acoustically rich representations appear to be linked with information about the lexical item in which the sound was encountered.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

After being stored, these phonetically rich representations contribute to an individual's production and perception. In perception, exemplars are activated based on their similarity to the incoming utterance; those that reach full activation fastest bias perception in their direction. During production, exemplars are activated based on a range of different kinds of information to which they are indexed; this information – stored at the time of encountering the utterance – includes phonetic, lexical, syntactic and pragmatic information as well as social information about the person who produced the utterance. Together, the activated exemplars influence the produced variant. Because production and perception are believed to rely on the same store of exemplars,[2] there is a production–perception loop whereby sounds perceived influence sounds produced, which are then perceived and stored (Pierrehumbert 2001).

While some exemplar-based models rely entirely on episodic memories at the exclusion of abstract representations (Hintzman 1986; Nosofsky 1986), we advocate one that includes episodic memories and abstract representations that are based on generalizations over those memories. Such a hybrid model would account for categorical effects as well as effects that are attributed to episodic memories (Pierrehumbert 2006). For mergers-in-progress, this means that speakers who have a single phoneme-level representation for two previously distinct sounds may nonetheless have somewhat separate, phonetically detailed distributions for the two sounds. These separate exemplar-level distributions come from exposure to people who maintain a distinction. Thus, people can have a single phoneme-level label and believe the sounds are the same but also have episodic memories where a distinction is represented.

Hay *et al*. (2010) present an overview of different studies on mergers, discussing whether some speech tasks involve activation of the acoustically rich exemplars while others activate abstract, phoneme-level representations. They argue that while the production and perception of real words activate exemplars, other tasks may require consultation of the phoneme level. When activation spreads across the different levels, it causes resonance (Johnson 2006: 495) – noise generated from the activation across the levels – the effect of which depends on whether a speaker has one or two categories for the sounds. For speakers who have two separate phoneme-level representations, resonance can increase the distinction between the two already distinct distributions of exemplars. For speakers who have a single phoneme-level representation, resonance increases the overlap between the distributions. In perception, listeners can rely on exemplars stored during interactions with speakers from dialects where the distinction is maintained.

To further test the predictions of the hybrid model, we have run two experiments that investigate differences between the production and perception of real and nonsense words. Nonsense words have a different set of predictions than real words because they

[2] While both production and perception rely on the same store of episodic memories, the exemplars that are activated during production are not necessarily (or even usually) the same as those activated during perception. For instance, people can perceive realizations of sounds that they themselves do not produce; they may have stored representations of these sounds that they use during perception, but these stored exemplars do not influence their production.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

have never been encountered before. The predictions are spelled out in more detail in the next section.

2.2 Predictions regarding lexical status for sounds undergoing merger

When discussing the status of a merger, it is conventional to declare a speaker 'merged' if they produce the sounds as acoustically identical. But in a hybrid model, there are different levels of merger to consider. A speaker might produce completely overlapping acoustic material, yet still – at some level – analyse the two phonemes as belonging to distinct abstract categories (based, perhaps, on their experience of other speakers producing the words). If both abstract categorization and episodic memories are jointly involved in production and perception, then the question of what abstract categories exist for a particular speaker/listener becomes centrally relevant.

In order to try to understand possible effects operating at different levels, the experiments reported in this article explore the behaviour of the production and perception of nonsense words. Nonsense words have not been encountered previously, so no episodic memories can be relied upon when producing or perceiving them. The use of nonsense words therefore forces operation at a more abstract level.

In production, if a speaker has the sounds categorized as belonging to different groups, we expect the nonsense words to be produced more distinctly than the real words, even if that speaker produces relatively merged real words. The very merged exemplar clouds would be bypassed because the lexical items do not exist at that level, and the speaker would instead produce something based on two different stored categories. If a speaker has both sounds collapsed into a single category, however, then there is no basis to expect that nonsense words will be any more distinct than real words. Instead, we might expect real words to be produced slightly more distinctly, as the speakers have encountered some distinct tokens to separate the distributions at the lexical level, but it is impossible to distinguish different tokens at the phoneme level.

Perception, however, holds different predictions. Almost all listeners have encountered at least some distinct tokens of real words from interacting with people who maintain a distinction and from the media. If perception involves matching incoming tokens to the best matching previously encountered episodes, real words would seem to have an inherent advantage in perception over nonsense words. The degree of this advantage should be affected by whether the speaker has one category or two. If they have one phonemic category and are forced to rely on an abstract representation when recognizing nonsense words, then nonsense words should be recognized no better than chance. If, however, they have two categories, then it should still be possible to recognize nonsense words with some accuracy. Nonsense words would then be recognized above chance, but not as well as real words.

The root of the difference in the direction of the predicted effects is that listeners are able to call on the extremes of their previous encounters (i.e. those that were produced distinctly). Therefore, in perception, exemplar clouds can help listeners detect quite subtle subphonemic differences. In production, however, speakers are likely to draw

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

Table 1. *Predictions of real and nonsense words, depending on the number of categories at the phoneme level*

	One category	Two categories
Production	Nonsense words should be as merged as real words – possibly even more.	Nonsense words should be more distinct than real words.
Perception	Nonsense words should be heard 'at chance' levels, and at lower levels than real words.	Nonsense words should be identified above chance, but with less accuracy than real words.

on episodes towards the centre of their distribution, and in line with their own identity and habitual motor patterns. At the word level, outlying 'distinct' exemplars are less likely to play an active role.

In vowels not undergoing merger, we expect the number of phonemic categories an individual has in production to match those in perception. In cases where merger appears near-complete, however, this might not follow. The presence of distinct categories in production needn't necessarily entail distinct categories in perception, nor vice versa. The predictions are outlined in table 1.

There is a lot of conjecture here, but the main thrust of the discussion is that we predict that nonsense words will behave quite differently from real words due to their lack of previously stored exemplars. We have tested this prediction in two mergers which are nearing completion: the ELLEN/ALLAN merger and the LOT/THOUGHT merger.

3 The ELLEN/ALLAN merger in New Zealand English

Our first experiment testing the differences between the effects of real and nonsense words on the production and perception of mergers took place in New Zealand (Thomas 2004; Thomas & Hay 2005). Thomas (2004) and Thomas & Hay (2005) reported the overall results of the study using relatively basic statistics. We here subject the results from experiment 1 to mixed-effects modelling, with particular attention to the potential role of word status in predicting production and/or perception.

In New Zealand English, there is a merger of DRESS and TRAP in prelateral position (Thomas & Hay 2005). The merger is nearly complete on TRAP; young males and females both produce the merger, and members of higher socioeconomic groups are least likely to produce it. While these vowels are distinct in most phonological contexts, most New Zealanders merge the vowels on TRAP prelaterally so that words such as *electricity* sound like *alectricity* to speakers of other dialects of English. This conditional merger creates ambiguity for a number of minimal pairs, such as *shall* and *shell*, *malady* and *melody*, and *salary* and *celery*.

It is not the subject of extensive public commentary, and there is no evidence that the merger is particularly salient to the community. It is not overtly stigmatized, nor – as far as we are aware – even noticed, by most speakers.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

For mergers that are conditioned by phonological environment, speakers have two distinct phoneme-level representations due to the environments where the distinction is maintained (e.g. *sad* and *said*). Therefore, New Zealanders have two distinct representations for DRESS and TRAP, despite their production of the merger before /l/. When producing real words that contain the vowels prelaterally, they can rely on their previously existing acoustically rich exemplars, which have overlapping distributions. When producing nonsense words, however, the speakers must rely on their (distinct) phoneme-level representations. While the categories may be collapsed prelaterally, the independently existing categories of DRESS and TRAP, together with the transparent orthography, probably help reinforce separate phonemic associations longer than in a non-conditioned merger. If we assume that they have two distinct phoneme realizations, subjects should be more merged when producing real words because they are relying on phonetically detailed word-based representations, and they should be less merged when producing nonsense words because they are relying on phoneme-based representations and the merger is only in some phonological contexts.

The experiment conducted to test these predictions compared behaviours across the production and perception of real and nonsense words (Thomas 2004; Thomas & Hay 2005). The nonsense words included both monosyllabic words (e.g. *chal*, *chel*, *val*, *vel*[3]), and bisyllabic words (e.g. *dallit*, *dellit*, *nallit, nellit*); the order of the production and perception task was blocked by word type (real words vs nonsense words) but not by syllabicity. The production data reported here come from isolated productions of the target words, which were produced together with a number of distracter filler items. Items were read from a word list.[4] The perception data comes from an identification task, in which the speech of a speaker who maintains the distinction was played. The target words were played in isolation in a pseudorandom order blocked by word type (real words vs nonsense words), and the participants circled one of two candidate answers on an answer sheet. Sixteen young New Zealanders took part in the task: 8 males and 8 females. A more detailed description of the methods can be found in Thomas (2004) and Thomas & Hay (2005).

3.1 ELLEN/ALLAN *production results*

F1 and F2 measurements were taken by hand, at a target point in the middle of the vowel's steady state. All participants showed a high degree of merger, roughly in the vicinity of their non-prelateral TRAP vowel. The high degree of overlap is shown in figures 1 and 2, for real and nonsense words. It can be seen that the vowel distributions of both types of words are highly overlapping. Overall, some small amount of separation exists for the nonsense words but not the real words.

[3] Some of the nonsense words may be nicknames and, thus, may have been encountered by some individuals. Such tokens, however, were avoided when possible.

[4] It is possible for the use of written words in the production task to have led to an effect of spelling.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

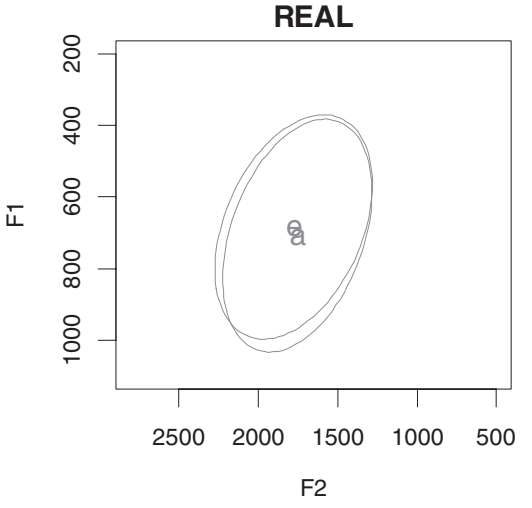


Figure 1. Ellipse plots of formant values at vowel targets for real words produced with ELLEN (e) and ALLAN (a) vowels

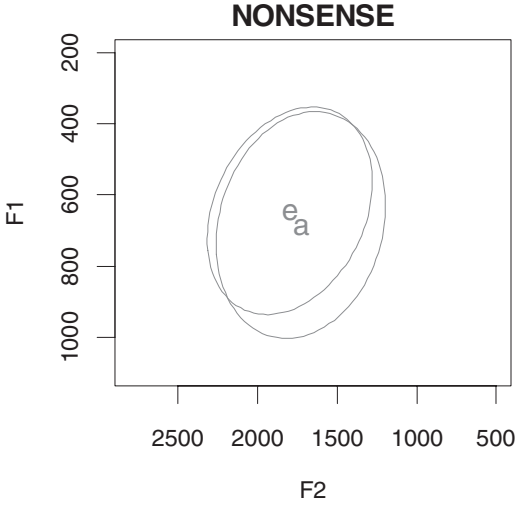


Figure 2. Ellipse plots of formant values at vowel targets for nonsense words produced with ELLEN (e) and ALLAN (a) vowels

In order to assess degree of merger in production, we quantified the degree of merger of each word pair for each speaker by calculating the Euclidean distance between the first two formants of the ELLEN token and the first two formants of the ALLAN token. We then modelled this Euclidean distance factor as a function of word type, testing interactions with social class and gender. A linear mixed-effects model was fit, with random effects for word and participant.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

Table 2. *Linear regression coefficients modelling Euclidean distance between* ELLEN/ALLAN *word pairs*

	Estimate	MCMC mean	HPD95 lower	HPD95 upper	pMCMC	Pr(>|t|)
(intercept)	257.05	256.94	189.776	322.996	0.001	0
Type = Real	−93.99	−93.91	135.025	−50.225	0.001	0
SES	−15.12	−15.06	−24.879	−4.723	0.01	0.0027
Type = Real : SES	11.35	11.36	5.479	17.61	0.001	0.0005

The social class index was established by assigning participants an index rating based on the occupations of their parents. The participants were students, so their own occupation could not be used to establish socioeconomic status. The Elley-Irving Scale (Elley & Irving 1985) was devised by New Zealand sociologists and provides occupations with a ranking, according to their prestige, as assessed by a combination of associated income and required education level. The scale runs from 1 (most prestigious) to 6. We established the Elley-Irving score for each parent, and then summed them to provide an estimate of socioeconomic status (SES) that runs from 2 to 12. There is a significant effect of social class, and of word status, and these interact with one another. The results are shown in table 2. The important columns to attend to here are the 'Estimate' column, which indicates the direction and magnitude of each effect, and the rightmost column, which shows the significance levels. In cases where there are interactions, the easiest way to conceptualize the model predictions is to plot the model estimates for the interaction, as shown in figure 3.

As shown by the slope of the line, speakers with a higher socioeconomic status (at the left of the graph) maintain a slightly greater distinction than those with a lower socioeconomic status. What can also be seen is that the largest difference is in the nonsense words.

What we suggest is that the slightly less merged speakers on the left are more likely to have two phonemic categories and to have the ELLEN/ALLAN words mapped to these categories. Because this is a conditioned merger, the people on the right also have a DRESS and TRAP category. But if they are very merged, this increases the chances that they may not use both categories for the prelateral environment. That is, they have both ELLEN and ALLAN associated with TRAP. If the *e*/*a* distinction does not exist at all in a speaker's phonological grammar, then when they produce nonsense words, they should be produced the same, regardless of how they are spelled (e.g. *sellit*, *sallit*). For such people, real words should be produced more distinctly than nonsense words because – at least for real words – the occasional distinct production has probably been encountered. However, if the speaker has their pre-/l/ forms accurately mapped to /e/ and /a/ categories (perhaps bootstrapped somewhat by spelling), then in nonsense words, these category representations are all that they have to rely on. The speaker should then

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

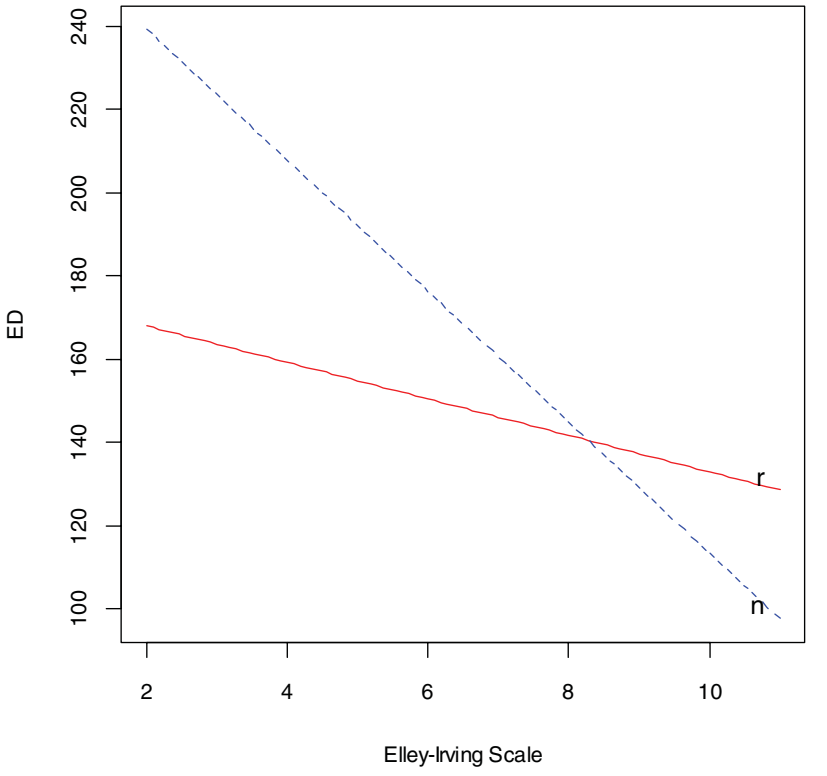


Figure 3. (Colour online) Predicted Euclidean distance between word pairs, as a function of the speaker's socioeconomic status and of the word's lexical status (r = real words, n = nonsense words). Lower numbers on the Elley-Irving Scale indicate a higher socioeconomic status

produce the nonsense words in a manner that is more distinct than their production of real words. An extra prediction of this model is that the variation across different levels of merger should be much more extreme for nonsense words than real words. For nonsense words, whether a speaker has one internal category or two has radically different consequences. For real words, the consequences are much more similar across the two scenarios. This is exactly the variation we see in figure 3; nonsense words are much more variable in their behaviour.

Looking at individuals' behaviour with real and nonsense words, shown in figure 4, we don't see a significant relationship between an individual's Euclidean distance in real words and their Euclidean distance in nonsense words. The graph reveals that many participants are near the x = y line; however, there are outliers. The two individuals in the top left of the graph both show a strong tendency to maximize distinction in nonsense words though they do not maintain much of a distinction in real words. Incidentally, these two speakers have the highest socioeconomic indices in the data set for females and males respectively.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

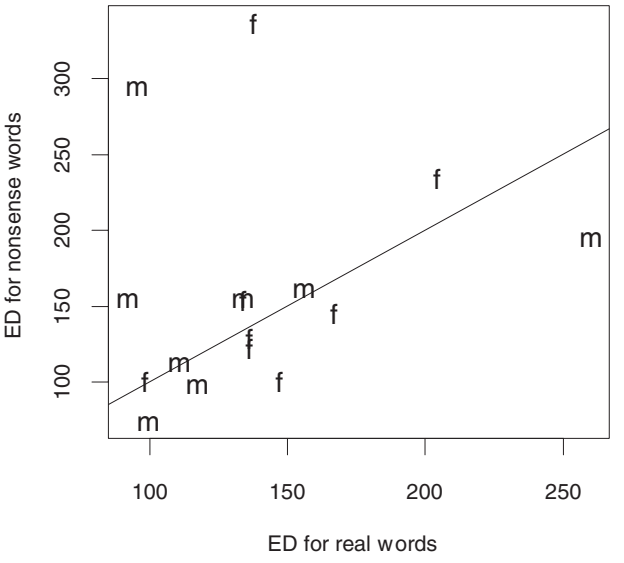


Figure 4. The relationship between individuals' production of real words and their production of nonsense words. The degree of distinction is quantified using Euclidean distance (ED) and is shown separately for males (m) and females (f)

3.2 ELLEN/ALLAN *perception results*

We now turn to the perception data, which is plotted separately for males and females, in figure 5. What is apparent is that listeners are for the most part operating above chance with respect to their perception of nonsense words. The fact that listeners can do this task with nonsense words to some degree indicates some level of abstract categorization of the two phonemes.

What can be seen is that accuracy in perception is remarkably high, especially given the high degree of merger in production. We also see an overall tendency for real words to be heard more accurately (as was predicted in table 1). The accuracy is much higher with /el/ which is to be expected; both auditorily and acoustically, the merger is much closer to the original /al/ than the original /el/ position. The [el] phonetic tokens are therefore not as ambiguous as the [al] phonetic tokens, which could more plausibly represent either category.

We fit a binary mixed-effects model to the data. Word type (real vs nonsense) was not significant. The significant factors in the model were the phoneme of the word being identified, in interaction with the parents' combined Elley-Irving score. Words containing /a/ were more likely to elicit an error, as noted above. Listeners of higher socioeconomic status made fewer mistakes, and this was particularly pronounced for /e/ words. The significant interaction from the model is shown in figure 6.

It is interesting that we don't observe an effect of word type. It is possible that this is simply because listeners are really quite good at this task, given their degree of mergedness. In figure 5, there is a visible trend for real words to be heard more

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

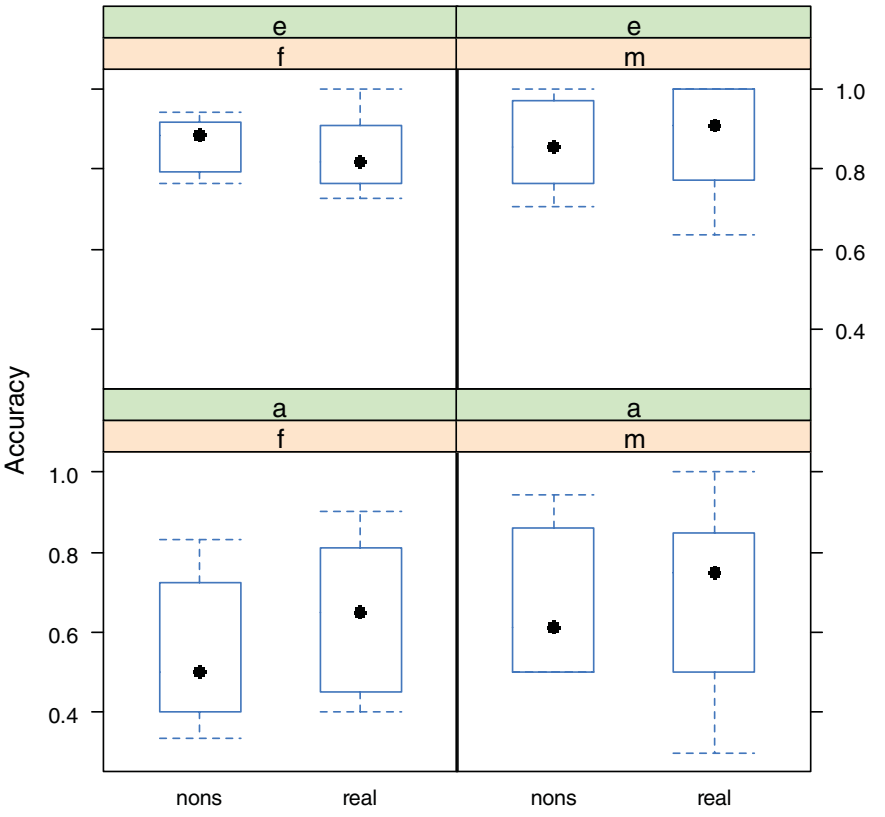


Figure 5. (Colour online) Accuracy in the perception task for females and males, plotted separately for /el/ words (top) and /al/ words (bottom)

accurately than nonsense words, and this trend is strongest in the pockets of the data with the lowest overall accuracy rates (see, e.g. female /a/ vowels in figure 5), lending some support to our interpretation of the lack of an effect being due to the high degree of accuracy on the task. While the tendency for participants to be more accurate with real words is not significant, it does trend in the predicted direction.

We observe a significant correlation between individuals' accuracy in real words and nonsense words (rho = .62, $p<.02$). This is shown in figure 7. The solid line shows the x = y line, and the dashed line shows a LOWESS scatterplot smoother (Cleveland 1981) fit through the data. The fact that the dashed line for the most part sits below x = y indicates that there is a slight general tendency for real words to be perceived more accurately than nonsense words (although, as we saw in the logistic regression, this tendency is not significant).

In production, our assumption is that speakers must rely on phoneme-level productions for nonsense words. Phonemic representations may also contain the non-prelateral (non-merged) tokens, which would pull the distributions apart. In perception,

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

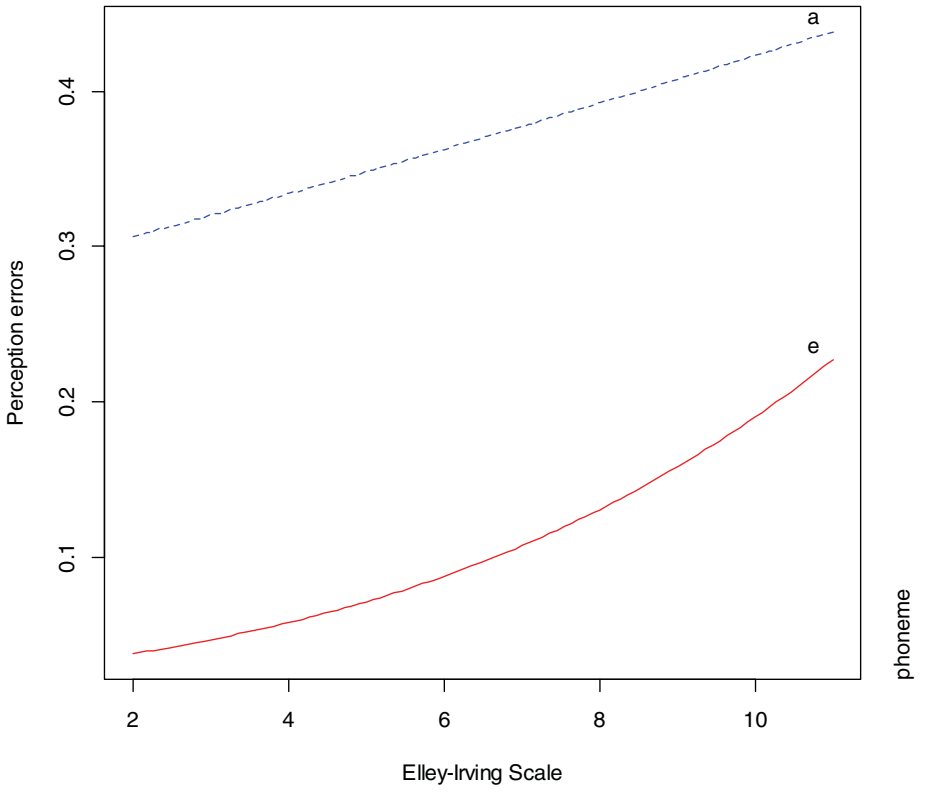


Figure 6. (Colour online) Interaction of vowel identity and a listener's socioeconomic status on likelihood of error. Lower numbers on the Elley-Irving indicate a higher socioeconomic status

the real-word exemplars include some distinct tokens, which help to identify the sounds (even though the listeners may feel like they are guessing). Thus, these results seem to provide evidence that individuals have both episodic phonetically detailed and abstract phoneme-based representations of sounds. The type of task (production or perception) and prior exposure to the word influence the extent to which an individual will rely on the phonetically detailed memories or the abstracted phonemic ones.

The interaction found with production for the ELLEN/ALLAN merger seems to suggest a difference between speakers who have two sounds mapped to a single phoneme category, versus those who have them mapped to two phoneme categories. The fact that it is a conditional merger probably strongly facilitates the two-phoneme analysis for speakers. Hence, most speakers maintain a greater distinction in nonsense words. It also increases the power of the phoneme-level analysis to increase a speaker's distinctiveness in production, as the non-prelateral forms of the vowels – part of the same analysis category – are fully distinct.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

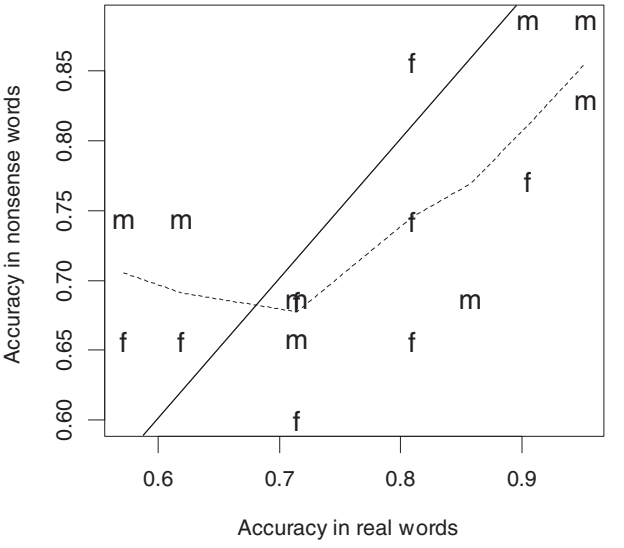


Figure 7. Correlation between accuracy at identifying real words and nonsense words during the perception task

3.3 Relationship between production and perception

If we assume that those speakers who have a bigger difference between nonsense words and real words in the production task are the speakers who are most likely to retain two independent phonemic categories, then we might expect that there is a relationship between the degree of nonsense-word bias in the production task, and the degree of nonsense-word bias displayed in the perception task. Or, at the very least, degree of distinction in production might be correlated with accuracy in the perception task, as one would guess that retention of multiple categories would aid identification in perception.

However, no such correlations hold. We can find no relationship between patterns of behaviour in the production and the perception task. Nor is there any relationship between an individual's average Euclidean distance in the production task, and their overall performance on the perception task. This is interesting, as it indicates that the relationship between production and perception in terms of how the merger is behaving is not a straightforward one. By way of example, figure 8 shows the non-connection between individuals' mean Euclidean distance in production and their accuracy in perception. There is no link.

To investigate this complex relationship further, the next experiment builds on results from the ELLEN/ALLAN experiment through exploring trends in production and perception within the context of a different merger found in a different dialect. Specifically, it investigates the production and perception of real and nonsense words containing the LOT and THOUGHT vowels, vowels that are undergoing a merger in

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

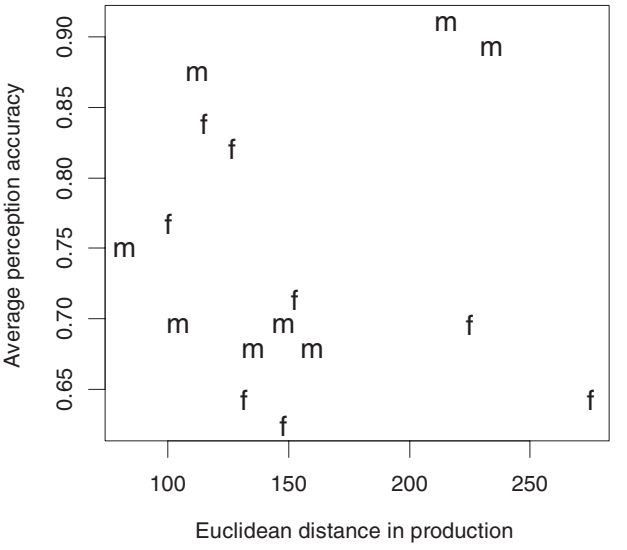


Figure 8. Non-correlation between production and perception

American English. In contrast with the ELLEN/ALLAN merger, we might expect that individuals who produce these mergers are less likely to have two distinct phoneme-level representations because this is not a conditional merger.

4 The LOT/THOUGHT merger in American English

The LOT/THOUGHT merger, also referred to as the *cot/caught* merger or the low back vowel merger, is the merger of /ɑ/ and /ɔ/. Speakers who have the merger do not distinguish between realizations of items in word pairs such as *cot–caught*, *collar–caller* and *pod–pawed.* The merger is found in the western contiguous United States (Clopper *et al.* 2005; Labov *et al.* 2006) and Hawai‘i, and it appears to be spreading to other areas as well, including parts of Missouri (Gordon 2006; Majors 2007), Kentucky (Irons 2007), and southern Illinois (Bigham 2010). The phonetic realization of the variant resulting from this merger varies depending on the region. The merger is resisted when followed by certain phonological environments for some speakers and regions, but other speakers merge these sounds in all contexts. There is little discussion in the literature of attitudes towards the merger, but many young speakers who produce the merger are unaware that the sounds are distinct in other dialects.

A follow-up study to the ELLEN/ALLAN experiment was conducted in Hawai‘i to determine how individuals from Hawai‘i and the western United States produce and perceive words differently depending on whether the word is real or not.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

4.1 *Experiment 2 methodology*

Twenty-three native speakers of English took part in the experiment, thirteen of whom were female. Sixteen were from Hawai'i. The other participants were from Colorado (2), Louisiana (1), Texas (1), Wisconsin (1), Iowa (1) and California (1). All participants were students at UH Manoa, and they received course credit for their participation. None of the participants had training in linguistics beyond the first-year level. Each participant was recorded reading 35 word pairs (14 nonsense, 21 real). The tokens were read in isolation rather than in pairs. Tokens were blocked by type (real word vs non-word) and the order was randomized within each block. The words varied in their following phonological context to determine whether some speakers maintained a distinction in some contexts but not in others. The following environments used were: /d/, /k/, /n/, /p/ and /t/. Fillers were used but were not analysed. Speakers were recorded in a quiet room at UH Manoa using a Tascam DR-07 portable digital audio recorder, and the recordings were made at 44.1kHz.

Using Praat, formant values were taken by hand at the vowel target: the middle of the steady state or, in the absence of a steady state, the point in the middle 80 per cent of the vowel's duration where the F1 and F2 values reach the target of the formant transitions.[5] In order to do the analysis, we calculated each participant's Euclidean distance between each member of the word pair. Thus, we included tokens in the data set only if the speaker produced both members of a pair; in some cases, participants misread a word, producing it with a completely different vowel (e.g. TRAP) so these were not included in the analysis. The analysed data set comprises 736 pairs of observations.

After recording the word pairs, participants took part in a binary, forced-choice identification task. This order of the tasks was consistent with the ELLEN/ALLAN experiment. The stimuli were recordings of distinct tokens produced by a female speaker from New York. Tokens were played in isolation, and participants indicated which of the two choices presented to them was the word they had heard. The task was blocked by word type. First, participants responded to 32 real words (16 pairs), and then they responded to 24 nonsense words (12 pairs). As with production, the words varied in their following phonological context. The order of stimuli was randomized in the production task and pseudorandomized in the perception task, so the order of the words and the following contexts was not the same across the two tasks. The order of the following phonological context varied between the real and nonsense word blocks but was the same across the different participants.

4.2 *LOT/THOUGHT production results*

In production, the speakers were very merged, regardless of the lexical status of the words. Ellipse plots for the nonsense words and real words are shown separately in

[5] The formant settings were as follows: 0.025 window length and 30 dB dynamic range. The maximum formant and number of formants used as settings varied by participant in order to maximize accuracy of the formant trace.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

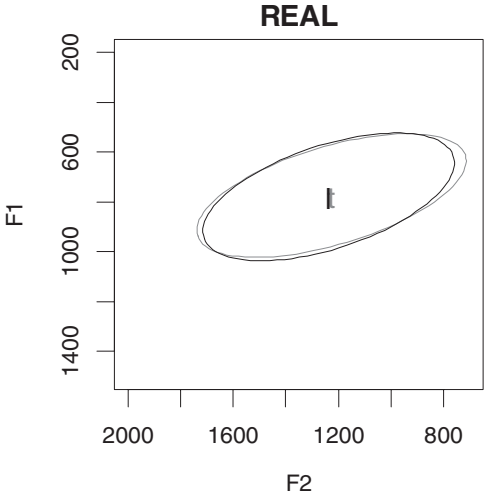


Figure 9. Ellipse plots for real words containing LOT (l) and THOUGHT (t)

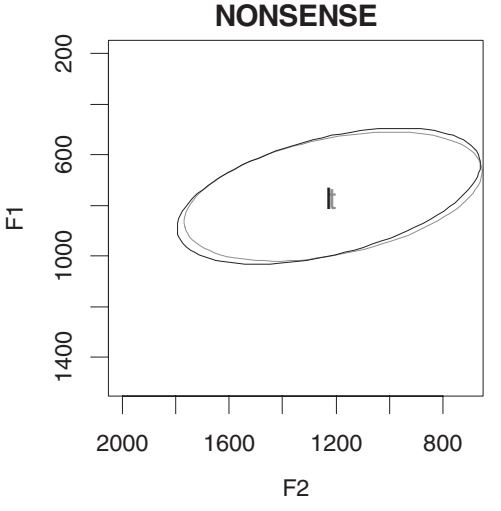


Figure 10. Ellipse plots for nonsense words containing LOT (l) and THOUGHT (t)

figures 9 and 10. Normalization for vocal tract size is unnecessary because pairs were matched for each speaker; fully merged realizations for all participants should produce ellipses that are completely overlapping, whereas the ellipses won't overlap entirely if the distinction is maintained by at least some participants. It can be seen from viewing the ellipses in figure 9 that both types of words were extremely merged.

We fit a linear mixed-effects model to the Euclidean distance, with speaker and word pair as random effects. The results are shown in table 3. The directions and magnitudes of effects are shown in the 'Estimate' column, with probability levels indicated in the

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

Table 3. *Linear mixed-effects model for Euclidean distance of* LOT/THOUGHT *word pairs*

	Estimate	MCMC mean	HPD95 lower	HPD95 upper	pMCMC	Pr(>|t|)
(intercept)	172.805	173.516	127.78	220.160	0.001	0.000
Males	−91.017	−91.945	−146.38	−22.53	0.002	0.0059
Real words	6.635	5.947	−29.84	47.82	0.770	0.7391
Non-alveolars	69.211	68.747	15.42	114.18	0.006	0.0093
Real words : non-alveolars	−115.080	−114.382	−180.70	−52.86	0.002	0.0010

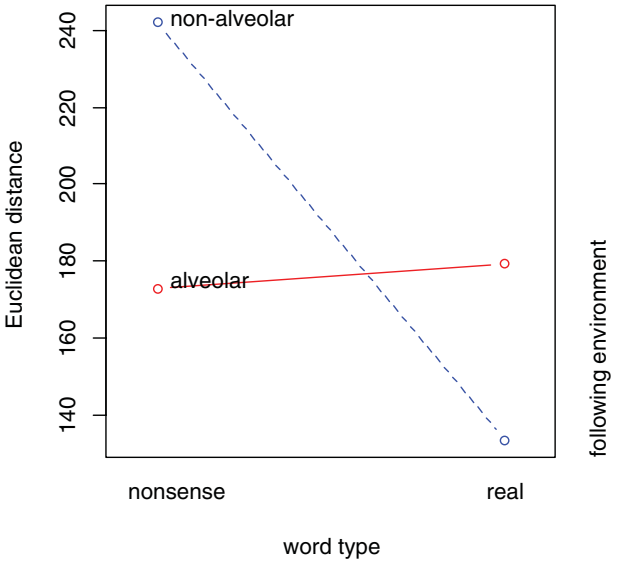


Figure 11. (Colour online) Interaction between lexical status and following phonological environment

rightmost column. We found an overall effect of gender, with males producing slightly but significantly lower Euclidean distances. This may be due to the different frequency ranges occupied by males and females, though no such effect was observed in the ELLEN/ALLAN data. We also found an effect of the lexical status of the word, which interacts with following environment. For vowels which are pre-alveolar, there is no effect of lexical status. However, for non-pre-alveolar vowels, there was a significant effect: nonsense words were produced with significantly greater distinction than real words. This interaction is seen in figure 11.

Our non-alveolar contexts are /k/ and /p/, both of which have a special status for this merger. The THOUGHT vowel is rarely followed by /p/; in fact, in our data we only have one real-word pair (*hop* and *Hawpe*) for this context. In contrast, there are

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

Table 4. *Binomial mixed-effects model for* LOT/THOUGHT *identification task*

	Estimate	Std. error	Z value	Pr(>|z|)
(intercept)	−0.9850	0.4671	−2.109	0.0350
Number	0.0232	0.0094	2.474	0.0134
Males	0.4513	0.2542	1.776	0.0758
Real words	0.7998	0.3245	2.464	0.0137
THOUGHT vowel	0.7317	0.1564	4.678	<0.0001
Gender∗ word type	−0.6274	0.2528	−2.482	0.0131

multiple words where THOUGHT is followed by /k/, but this environment resists the merger more than any other environment (Labov *et al.* 2006: 64). Interestingly, the speakers in our experiment actually produce less of a distinction in this context in real words than they do in other contexts. This result is interpretable if we assume that there are distinct subphonemic categories for the vowels in this context. Because this context resists the merger, the speakers in our study are likely to have encountered distinct tokens and, therefore, likely to have distinct exemplar-level representations for this phonological environment. During the production of real words in this task, our speakers rely on exemplars that index their habitual articulations, producing merged tokens. When producing nonsense words, however, they must rely on an abstraction, one which was generated based on exemplars from the extremes of their distributions. These extremes are tokens produced by speakers who maintain a distinction in this context but not other contexts. The fact that the real-word status does not affect the vowels in the same way across the different phonological contexts may indicate that there is a level of abstraction which contains the entire rhyme (i.e. is not just a single phoneme).

For /p/ it is difficult to say because few studies examine the merger in this context. Its behaviour in our data is similar to /k/, but there is only one real-word word pair for /p/ and the THOUGHT word in this pair is a proper name. Therefore, this trend should be viewed with some caution; the effect may not be related to alveolars versus non-alveolars but instead /k/ versus other sounds.

While the rate of distinction maintained is overall larger for nonsense words (as driven by the non-alveolar pairs), there is a strong correlation between individuals' behaviour in real words and nonsense words. Figure 12 plots individuals' mean Euclidean distance maintained in real-word pairs against the mean Euclidean distance produced in nonsense word pairs. The correlation is highly robust (rho = .76., p < .0001). The more of a distinction an individual maintains in their production of real words, the more of a distinction they are likely to retain in their nonsense words, and vice versa. The solid line shows an x = y line fit through the data. The dashed line shows a LOWESS scatterplot smoother fit through the data. It can be seen, both by eyeballing the points and by observing the scatterplot line, that there is a slight overall

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

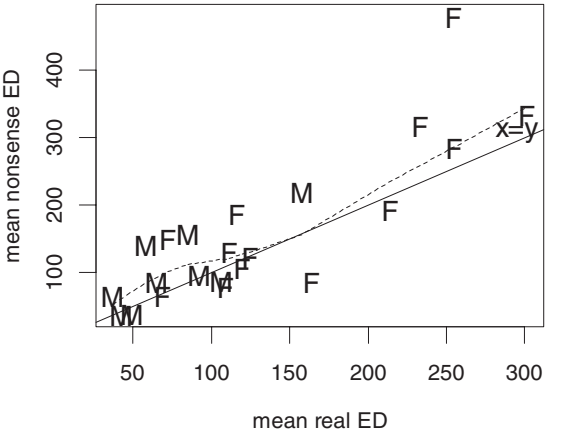


Figure 12. Correlation between degree of distinction in real and nonsense word pairs. Larger numbers on the x and y axes indicate greater distinction. The solid line shows the x-y line and the dotted line shows the LOWESS scatterplot smoother fit to the data.

tendency for individuals to maintain more of a distinction in nonsense words than real words.

4.3 LOT/THOUGHT perception results

Figure 13 shows boxplots of the average accuracy rates, calculated by speaker. With the exception of female responses to nonsense LOT words, performance in all cases was reasonably accurate, especially given the high degree of merger that was observed in production. Responses to the real words are fairly well matched across males and females. The responses to the nonsense words, however, differ. As with the ELLEN/ALLAN merger, accuracy is lowest with the variant on which the sounds have merged (here, LOT).

We fit a binary mixed-effects model with accuracy as the dependent variable and participant and word as random effects. This is shown in table 4. There was an overall effect of order of presentation, with accuracy increasing throughout the experiment. We assume this is an effect of training, aided by spelling. Including order of presentation (Number) in the model allows us to control statistically for this effect while examining effects from other factors. As shown in the model, THOUGHT vowels were identified more accurately than LOT vowels, and there was a significant interaction between gender and word type. Though participants responded to the real words before the nonsense words, real words were heard more accurately than nonsense words and this was particularly true for women. This interaction is shown in figure 14.

Real words are perceived significantly more accurately than nonsense words, and this result is predicted in table 1 and matches the (non-significant) trend in experiment 1. It is interesting that, despite being completely merged in production, the females are

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

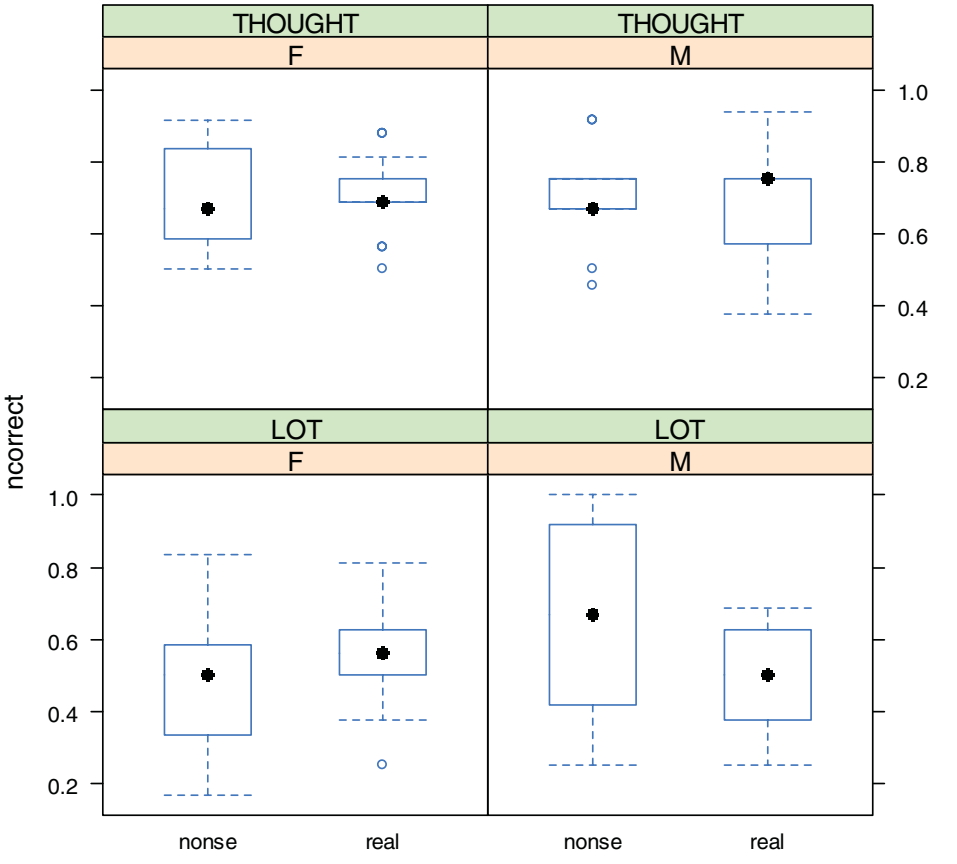


Figure 13. (Colour online) Accuracy on the identification task for real and nonsense words, and females (F) and males (M). Accuracy for THOUGHT tokens is shown in the top two plots, and for LOT tokens in the bottom two plots

more accurate with real words and less accurate with nonsense words. This interaction is a good illustration of why work with real words may not be revealing the full story, and why nonsense words may provide valuable arsenal in the merger-detective's toolkit.

For perception, there is no overall correlation between accuracy rates in real and nonsense words. The (non-)relationship is shown in figure 15. The solid line shows the $x = y$ line. The significant interaction between lexical status and gender is somewhat evident on the graph, with most of the males falling above the line, being more accurate with nonsense words. There is a cluster of females, however, who are reasonably accurate with real words, but performing around chance with nonsense words. Note that no one falls into the opposite pattern; there is no one who is highly accurate with nonsense words, yet performs at chance with real words. (This would be the top left quadrant of the graph.)

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

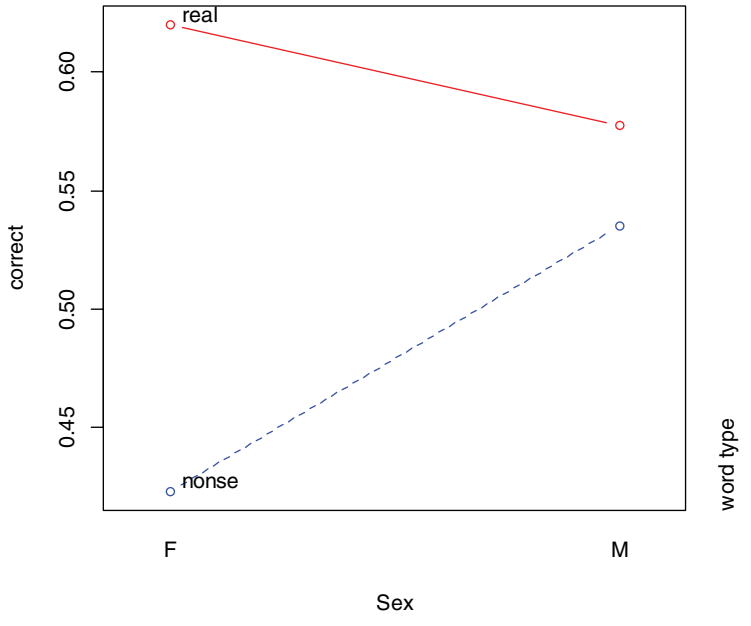


Figure 14. (Colour online) Interaction between speaker sex and lexical status in predicting accuracy during the identification task

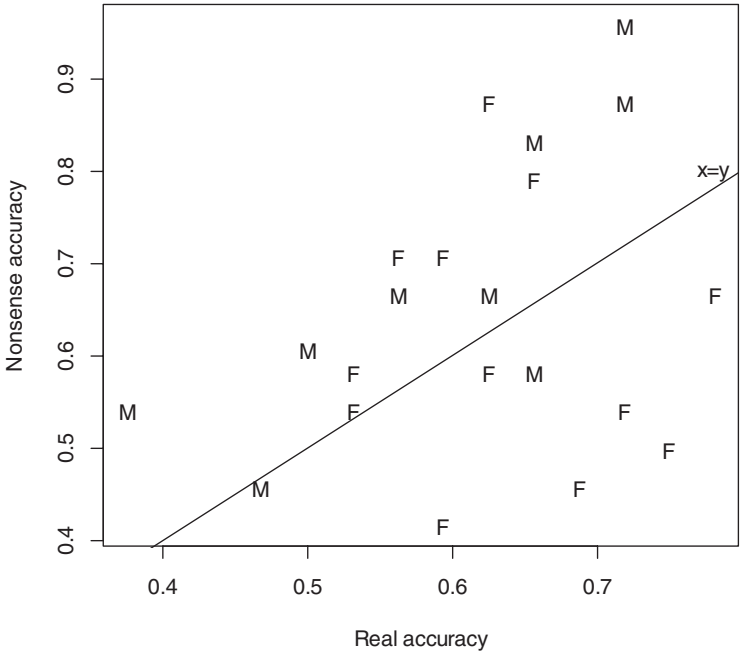


Figure 15. Non-correlation between accuracy of real and nonsense words. The solid line is the x–y line.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

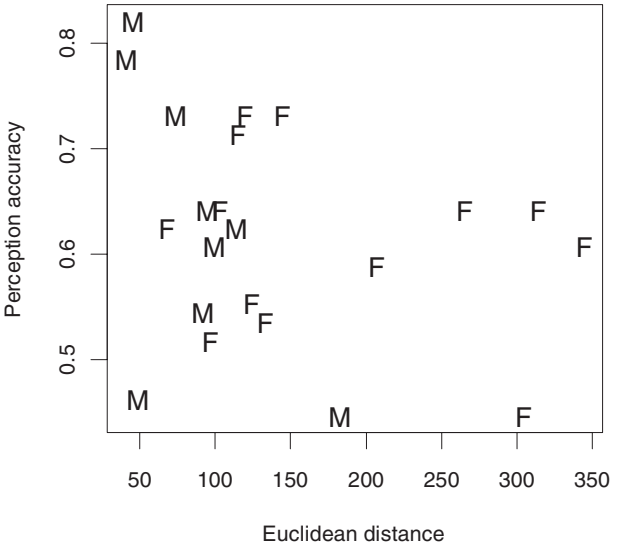


Figure 16. Individuals' mean Euclidean distance, and their accuracy on the LOT/THOUGHT perception task

Interestingly, when fit separately, the males show a significant relationship between their perception of real words and non-words (rho .81, $p<.01$). This relationship is absent for the females.

4.4 Relationship between production and perception

We have explored the possibility of a correlation between individuals' behaviour in production and perception. There is no correlation between degree of distinction in production and accuracy in perception, either overall, or considered separately for real and nonsense words. By way of illustration, figure 16 shows the apparently completely random association between individuals' mean Euclidean distance, and their accuracy on the perception task.

5 Discussion

This exploration of the effect of lexical status in mergers is largely consistent with the predictions laid out in table 1. There are a number of important points to note:

(1) Nonsense words and real words are different.

Overall, we find evidence that real words and nonsense words behave differently from one another. This points to the value of nonsense words in terms of trying to come to grips both with the characteristics of individual mergers, and also with understanding mechanisms of merger more generally. It also provides general evidence in favour of

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

some kind of hybrid model of speech production and perception, and provides evidence that some kind of word-level exemplar store is generally involved for real words.

(2) Nonsense words can be distinguished to some degree in perception.
In both sets of data, listeners are performing surprisingly well at identifying nonsense words. Despite almost complete overlap of the vowel spaces in both mergers, many people show reasonable accuracy with perception of nonsense words. This indicates that, at some level, these listeners still have the phonemes/word classes accurately generalized into separate categories. We have shown in our previous work on NEAR/SQUARE that listeners displaying mergers in production can nonetheless do very well in perception tasks, even when they feel that they are guessing (Hay *et al.* 2006b). In that work we argued that listeners were relying on stored exemplars from distinct speakers in order to perform the task. Here we show that this ability seems to extend in some cases even to nonsense words, which requires some level of generalization to explain.

(3) Nonsense words are less accurately identified than real words in perception.
This result was predicted based on the assumption that, for real words, listeners can rely on previously experienced, outlying, distinct exemplars. The effect is significant for LOT/THOUGHT. It is not significant for ELLEN/ALLAN, though it does trend in the expected direction.

(4) Nonsense words are more distinct than real words in production.
This was predicted based on the assumption that if a speaker has two distinct categories generalized, then forcing them to rely on these categories might elicit more distinct productions. We might especially expect this to be true in the case of the conditioned merger, where non-prelateral /e/ and /a/ may also form part of these categories, thereby forcing the distributions apart. In the ELLEN/ALLAN data, we observed the expected effect for the most distinct speakers (i.e. those of higher socioeconomic status). We argue that these are the speakers who are most likely to retain separate categories. The effect is also present for the LOT/THOUGHT merger in the predicted direction, but only when the sound is not followed by an alveolar.

(5) There is a relationship between nonsense words and real words, for both production and perception.
Taken as a whole, the results seem to suggest a reasonably strong relationship between real and nonsense words within production and perception. If you tend to make a distinction for real words, you are more likely to make a distinction for nonsense words. This is seen much more strongly in some pockets of the data than others. We see it most strongly in the LOT/THOUGHT production data (figure 12), where the relationship between production of nonsense words and real words is very tight. While there does seem to be this relationship for many speakers in the ELLEN/ALLAN data, this is not a significant correlation. For ELLEN/ALLAN perception, there is a significant correlation between accuracy in real words and accuracy for non-words. In LOT/THOUGHT perception the same trend is significant for the males but not the females. But in both sets of data we have an empty graph quadrant: there are no cases of speakers who are bad at perceiving real words, but good at perceiving nonsense words. It is not clear whether much can be read into the different levels of strength of these associations,

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

but looked at overall, there is reasonable evidence that the real words and nonsense words are not behaving completely independently of one another. This would make sense as we assume that any categorizations that speakers have are generalizations over existing words. That is, the abstractions used to produce/perceive the nonsense words are not generated independently of an individual's experience of real words.

(6) In neither group is there any relationship between production and perception. Somewhat surprisingly, we were not able to find any relationship between an individual's behaviour in production and their behaviour in perception. There is no significant correlation for either data set. We don't believe these two modalities operate completely independently of one another. Indeed, the entire premise of a model which includes an episodic word store is that this evolves via a production–perception loop, where production and perception operate on the same episodic memories. That said, production and perception are very different from one another in terms of the range of the episodic distributions generally available to the individual. Individuals may use exemplars towards the extremes of their distributions in perception, if that is what the incoming speech signal best matches. But they are much less likely to produce a token that comes from the extremes of their distribution. Because most of the participants in our studies are quite merged and probably interact mostly with people who are quite merged, the speaker they are listening to in the perception task is in many ways atypical, representing a different dialect. The listener may well have access to this dialect in perception, but – due to their lower frequency – memories of previous encounters with this dialect are not overly likely to be influential in production. This one element most likely decouples overall production and perception more than one might otherwise expect.

A reviewer points out that there may also be task elements that are forcing a greater decoupling here than might really exist. For example, it is possible that participants who do not normally make a distinction may be forced to concentrate more in the perception task – and rely more on the extremes of their experience – than participants who are less merged.

The fact that we see no relationship between the effects of lexical status in production and perception is also interesting; if a listener has much more distinct nonsense words than real words in production, this predicts nothing about whether they will be more accurate with real or nonsense words in perception. This raises the possibility that categorization into two abstractions for perception purposes might not necessarily imply the same is true for production, and vice versa. While we believe that the episodic word store is shared between production and perception, there is no compelling reason why more abstract levels of categorization might not exist separately for the different modalities. Indeed, it might make sense that there are different spheres over which a speaker/listener generalizes in order to make abstractions associated with production and perception. Production may be generalized over a more reduced set, as it's more likely to be influenced by one's own exemplars and habitual motor patterns. Perceptual categories, on the other hand, may need to be more elastic, in order to understand a wide range of speakers.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

6 Conclusion

This exploration into real words and nonsense words in production and perception of mergers-in-progress suggests that this is a fruitful line of research. If multiple levels of representation are involved in production and perception, then we need innovative ways of trying to understand how these levels – and the relationships between the levels – work. The tasks presented here constitute a very simple, initial exploration of the idea that manipulating lexical status might bear some fruit. The parallels between the two data sets in terms of production and perception, and the relationship between the two, provide some solid groundwork for building up a richer understanding of the dynamics of speech perception and production in the context of categorical and conditioned merger.

Authors' addresses:
University of Canterbury
Private Bag 4800
Christchurch
New Zealand
jen.hay@canterbury.ac.nz

University of Hawai'i at Mānoa
Department of Linguistics
561 Moore Hall
1890 East-West Road
Honolulu
Hawai'i 96822
kdrager@hawaii.edu

United Arab Emirates University
PO Box 17172
Al Ain, Abu Dhabi
United Arab Emirates
b.thomas@uaeu.ac.ae

References

Aylett, M. & A. Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence and duration in spontaneous speech. *Language and Speech* 47(1), 31–56.

Bailey, Guy, Tom Wikle, Jan Tillery & Lori Sand. 1993. Some patterns of linguistic diffusion. *Language Variation and Change* 5, 359–90.

Baker, Rachel E. & Ann Bradlow. 2009. Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech* 52(4), 391–413.

Bauer, Laurie. 1986. Notes on New Zealand English phonetics and phonology. *English Worldwide* 7, 225–58.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

Bigham, Douglas S. 2010. Correlation of the low-back vowel merger and TRAP-retraction. *University of Pennsylvania Working Papers in Linguistics* 15(2), 21–31.
Boberg, Charles. 2000. Geolinguistic diffusion and the US–Canada border. *Language Variation and Change* 12, 1–24.
Cleveland, William S. 1981. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician* 35(1), 54.
Clopper, Cynthia G., David B. Pisoni & Kenneth de Jong. 2005. Acoustic characteristics of the vowel systems of sex regional varieties of American English. *Journal of the Acoustical Society of America* 118(3), 1661–76.
Drager, Katie K. 2010. Sensitivity to grammatical and sociophonetic variability in perception. *Laboratory Phonology* 1(1), 93–120.
Drager, Katie K. 2011a. Sociophonetic variation and the lemma. *Journal of Phonetics* 39(4), 694–707.
Drager, Katie K. 2011b. Speaker age and vowel perception. *Language and Speech* 54(1), 99–121.
Drager, Katie K., Rebecca Clifford & Jennifer Hay. 2011. The production and perception of a low back vowel merger. Paper presented at New Ways of Analyzing Variation 40. Georgetown, October 2011.
Elley, W. B. & J. C. Irving. 1985. The Elley-Irving socio-economic index: 1981 census revision. *New Zealand Journal of Educational Studies* 20, 115–28.
Foulkes, Paul, Gerard Docherty, Ghada Khattab & Malcah Yaeger-Dror. 2010. Sound judgments: Perception of indexical features in children's speech. In Dennis R. Preston & Nancy Niedzielski (eds.), *A reader in sociophonetics*. New York: De Gruyter.
Gahl, S. 2009. 'Time' and 'thyme' are not homophones: The effect of lemma frequency on word durations in a corpus of spontaneous speech. *Language* 84(3), 474–96.
Goldinger, Stephen D. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(5), 1166–83.
Gordon, Elizabeth & Margaret A. Maclagan. 2001. 'Capturing a sound change': A real time study over 15 years of the NEAR/SQUARE diphthong merger in New Zealand English. *Australian Journal of Linguistics* 21(2), 215–38.
Gordon, Matthew J. 2006. Tracking the low back merger in Missouri. In Thomas Edward Murray & Beth Lee Simon (eds.), *Language variation and change in the American Midland: A new look at 'Heartland' English*. Philadelphia: John Benjamins, 57–68.
Hay, Jennifer & Joan Bresnan. 2006. Spoken syntax: The phonetics of *giving a hand* in New Zealand English. *The Linguistic Review* 23, 321–49.
Hay, Jennifer, Katie Drager & Paul Warren. 2009. Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. *Australian Journal of Linguistics* 29(2), 269–85.
Hay, Jennifer, Katie Drager & Paul Warren. 2010. Short-term exposure to one dialect affects processing of another. *Language and Speech* 53(4), 447–71.
Hay, Jennifer, Aaron Nolan & Katie Drager. 2006a. From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review* 23, 351–79.
Hay, Jennifer, Paul Warren & Katie Drager. 2006b. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34(4), 458–84.
Hintzman, Douglas L. 1986. 'Schema abstraction' in a multiple-trace memory model. *Psychological Review* 93(4), 411–28.
Holmes, Janet & Allan Bell. 1992. On shear markets and sharing sheep: The merger of EAR and AIR diphthongs in New Zealand English. *Language Variation and Change* 4, 251–73.
Irons, Terry Lynn. 2007. On the status of the low back vowels in Kentucky English: More evidence of merger. *Language Variation and Change* 19, 137–80.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (eds.), *Talker variability in speech processing*, 145–65. San Diego: Academic Press.
Johnson, Keith. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34(4), 485–99.
Jurafsky, Daniel, Alan Bell & Cynthia Girand. 2002. The role of the lemma in form variation. In Carlos Gussenhoven & Natasha Warner (eds.), *Papers in laboratory phonology VII*, 1–34. Berlin and New York: Mouton de Gruyter.
Labov, William. 2001. *Principles of linguistic change*, vol. 2: *Social factors*. Oxford: Blackwell.
Labov, William. 2010. *Principles of linguistic change*, vol. 3: *Cognitive and cultural factors.* Oxford: Wiley Blackwell.
Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology* and *sound change*. Berlin: Mouton de Gruyter.
Labov, William, Mark Karan & Corey Miller. 1991. Near mergers and the suspension of phonemic contrast. *Language Variation and Change* 3, 33–74.
Maclagan, Margaret & Elizabeth Gordon. 1996. Out of the AIR and into the EAR: Another view of the New Zealand diphthong merger. *Language Variation and Change* 8, 125–47.
Majors, Tivoli. 2007. Low back vowel merger in Missouri speech: Acoustic description and explanation. *American Speech* 80(2), 165–79.
Nielsen, Kuniko. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 38: 132–42.
Nosofsky, R. M. 1986. Attention, similarity, and identification-categorization relationship. *Journal of Experimental Psychology* 115, 39–57.
Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. J. Hopper (eds.), *Frequency effects and emergent grammar*, 137–58. Amsterdam: John Benjamins.
Pierrehumbert, Janet B. 2006. The next toolkit. *Journal of Phonetics* 34(4), 516–30.
Strand, Elizabeth A. & Keith Johnson. 1996. Gradient and visual speaker normalization in the perception of fricatives. In Dafydd Gibbon (ed.), *Natural language processing and speech technology*, 14–26. Berlin: Mouton de Gruyter.
Thomas, Brynmor. 2004. In support of an exemplar-based approach to speech perception and production: A case study on the merging of pre-lateral DRESS and TRAP in New Zealand English. MA thesis, University of Canterbury.
Thomas, Brynmor & Jennifer Hay. 2005. A pleasant malady: The ELLEN/ALLAN merger in New Zealand English. *Te Reo* 48, 69–93.
Ullman, T. Michael, Ivy V. Estabrooke, Karsten Steinhauer, Claudia Brovetto, Roumyana Pancheva, Kaori Ozawa, Kristen Mordecai & Pauline Maki. 2002. Sex differences in the neurocognition of language. *Brain and Language* 83, 141–3.
Wells, John C. 1982. *Accents of English*, 3 volumes. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press

Appendix

Table A1. *Words used in the ELLEN/ALLAN perception task*

1-syll nons	2-syll nons	real
del-dal	dellit-dallit	celery-salary
kel-kal	kellit-kallit	elf-alf
lel-lal	lellit-lallit	ellen-allan
mel-mal	mellit-mallit	ellie-alley
nel-nal	nellit-nallit	kelvin-calvin
rel-ral	rellit-rallit	mellow-mallow
sel-sal	sellit-sallit	melody-malady
tel-tal	tellit-tallit	pellet-palate
zel-zal	zellit-zallit	shell-shall
		telly-tally

Additional words included in the production (but not perception) task were: *alligator*, *elevator*, *mallet*, *chel*, *chal*, *fel*, *fal*, *gel, gal*, *jel*, *jal*, *vel*, *val*, *vellit*, *vallit, fellit*, *fallit*, *gellit*, *gallit*, *jellit*, *jallit*.

Table A2. *Word pairs used as stimuli in the LOT/THOUGHT production task*

	d	t	n	k	p
real	cod-cawed nod-gnawed pod-pawed sod-sawed	bot-bought cot-caught knot-naught Ott-ought sot-sought tot-taught rot-wrought	Don-dawn fond-fawned pond-pawned Von-Vaughn yon-yawn	fox-Fawkes hock-hawk stock-stalk tock-talk wok-walk	hop-Hawpe
nonsense	dodd-dawd fod-fawd	chot-chawt drot-drawt vot-vaught	stonn-stawn tonn-tawn zon-zawn	fock-fawk grock-grawk vock-vawk	gopp-gawp ropp-rawp

Table A3. *Word pairs used as stimuli in the LOT/THOUGHT perception task*

	d	t	n	k	p
real	pod-pawed cod-cawed sod-sawed nod-gnawed	cot-caught tot-taught knot-naught rot-wrought	yon-yawn don-dawn pond-pawned fond-fawned	wok-walk hock-hawk tock-talk stock-stalk	
nonsense	dodd-dawd vod-vawd fod-fawd	zot-zaught vot-vaught drot-drawt mot-maught chot-chawt		vock-vawk	dop-dawp gopp-gawp ropp-rawp

https://doi.org/10.1017/S1360674313000026 Published online by Cambridge University Press