Variability in the pronunciation of non-native English *the*: Effects of frequency and disfluencies

JESSAMYN SCHERTZ AND MIRJAM ERNESTUS

Abstract

This study examines how lexical frequency and planning problems can predict phonetic variability in the function word 'the' in conversational speech produced by non-native speakers of English. We examined 3180 tokens of 'the' drawn from English conversations between native speakers of Czech or Norwegian. Using regression models, we investigated the effect of following word frequency and disfluencies on three phonetic parameters: vowel duration, vowel quality, and consonant quality. Overall, the non-native speakers showed variation that is very similar to the variation displayed by native speakers of English. Like native speakers, Czech speakers showed an effect of frequency on vowel durations, which were shorter in more frequent word sequences. Both groups of speakers showed an effect of frequency on consonant quality: the substitution of another consonant for /ð/ occurred more often in the context of more frequent words. The speakers in this study also showed a native-like allophonic distinction in vowel quality, in which /ði/ occurs more often before vowels and /ðə/ before consonants. Vowel durations were longer in the presence of following disfluencies, again mirroring patterns in native speakers, and the consonant quality was more likely to be the target /ð/ before disfluencies, as opposed to a different consonant. The fact that non-native speakers show native-like sensitivity to lexical frequency and disfluencies suggests that these effects are consequences of a general, non-language-specific production mechanism governing language planning. On the other hand, the non-native speakers in this study did not show native-like patterns of vowel quality in the presence of disfluencies, suggesting that the pattern attested in native speakers of English may result from language-specific processes separate from the general production mechanisms.

Keywords: *pronunciation variation, non-native speech, phonetics, lexical probability, disfluencies*

1. Introduction

Recent work has investigated the realization and the predictors of 'acoustic reduction,' the lenition or omission of segments especially prevalent in casual speech (see Ernestus and Warner (2011) for an overview). For example, Johnson (2004) found that in a corpus of American English, 25% of word tokens exhibit omission of at least one segment, while entire syllables are missing from the acoustic signal for 6% of word tokens. Furthermore, these studies show that acoustic reduction tends to occur to a greater extent in more frequent syllables, words, or word sequences (e.g. Aylett and Turk 2004, Pluymaekers et al. 2005, Bell et al. 2009). On the other hand, words in disfluent contexts within conversations are more likely to be pronounced in fuller forms with longer durations (Bell et al. 2003). Work on these topics has focused almost exclusively on native speech, leaving open the question of whether the same predictors of reduction and clear speech hold for non-native speakers, and if so, if reduction is realized in the same way.

Such effects of frequency and disfluencies may be indicators of speech planning: if more frequent words are more easily accessible in the mental lexicon (e.g. Jescheniak and Levelt 1994), speakers are able to continue speaking at a fast speech rate (resulting in more reduction), while disfluencies are a symptom of planning problems, causing speakers to slow down (resulting in less reduction). Under the assumption that there is a general production mechanism governing speech planning, non-natives would be expected to show similar effects. A rare example of work examining reduction in non-native speech found that Korean and Chinese speakers of English show similar effects of word frequency on word duration as native speakers, though to a lesser extent (Baker et al. 2011). This study focused on read speech, so the question remains whether these effects are also present in non-native speakers in more spontaneous conversational settings. Furthermore, it might then be expected that non-native speakers, like native speakers, show fuller forms in disfluent contexts.

An additional point of interest is the realization of segments not present in the speakers' native languages. In particular, what are the 'reduced' and 'full' forms of segments that may be difficult for the speakers to pronounce, and how do these realizations compare to those of native speakers?

In order to investigate these questions, we examined the pronunciation of the function word *the* in conversational English spoken by Czech and Norwegian speakers. Norwegian and Czech make for an interesting comparison because they differ in how closely they are related to English, with Norwegian being more related and Czech being less related. Furthermore, we had access to two corpora drawn from the same task. We chose to focus on a single function word in order to isolate the sources of variability in the most controlled way possible. Furthermore, *the* includes the interdental fricative /ð/, a sound which is present in neither Czech nor Norwegian, allowing us to observe the variation in phonetic realizations of a notoriously difficult sound for non-native speakers. We measured variability in terms of three factors: vowel duration, vowel quality, and consonant quality.

If non-native speech shows patterns similar to those of native speakers, we expect vowels to be shorter and more reduced in the context of more frequent word combinations, while the opposite is expected in the case of disfluencies. We might also expect the vowel /i/ to be the signal of a disfluency, as has been found for native speakers (Fox Tree and Clark 1997). For consonant quality, preliminary observations in the corpus suggest that substitution by a different phone (such as [d] or [ɾ]) is very common. This variation may mirror patterns in native speakers, who often substitute other consonants for /ð/ (e.g. Manuel 1995, Cao 2002, Zhao 2007, 2010). So far, no study has provided information about whether the type and rate of substitution is affected by lexical frequency or disfluency, as we have in the present study. Since /ð/ is a notoriously difficult sound for non-native speakers, we expect them to "reduce" their pronunciations by choosing an easier segment. Therefore, we predict that non-native speakers will produce more deviant segments in contexts favoring reduction.

In Study 1, we examined the effect of the frequency of the word following the target *the* in fluent non-native speech. We looked at variation in vowel duration, vowel quality, and consonant quality in order to compare the observed varation in vowel duration and quality with previous findings for native speakers, and in order to explore the effect of frequency on the realization of consonants in non-native speakers. In Study 2, we turned to the effects of disfluencies on these same three phonetic variables to see if non-native speakers use fuller forms in the context of disfluencies, as native speakers do.

2. Study 1: Variability of *the* in fluent non-native English

2.1. *Method*

The data used in this study were drawn from the the Kachna Corpus (Spilková et al. 2010), which contains English conversations between pairs of native speakers of either Czech or Norwegian. One speaker is given a cartoon drawing and describes it to the other speaker, who is asked to try to replicate the picture without seeing it. The Norwegian speakers (six males and four females) recruited for the task were all university students, while the Czech speakers (five males and five females) were university students of English or employees in organizations where English is used on a regular basis. No standardized measures of proficiency were collected for the corpus, but a high level of English proficiency is evident in the recordings: all speakers were able to complete the 30-60-minute conversational task without apparent effort or fatigue, and vocabulary problems were rare, usually resulting from less common objects that appeared in the pictures (e.g. “shovel”). There is no obvious difference in the level of English between the two language groups. If there are subtle proficiency differences between speakers, those that are relevant to our findings should emerge in our statistical analysis.

In order to obtain information about phone-level variation, we automatically created a broad phonetic transcription, which was aligned with the speech signal, using the Hidden Markov Model toolkit HTK (Young et al. 2002). First, the orthographically transcribed conversations, comprising a total of approximately 7.5 hours of speech and 20 speakers, were split into utterances delineated by speaker changes or silences of over 200 ms. Portions of conversation containing overlapping speech or background noise were omitted from the transcription process, as were the few utterances containing non-English words or mispronunciations suggesting that the target pronunciation was not the canonical English pronunciation (such as [bɹowd] for *broad*), as this would have posed problems for the automatic segmenter. The remaining utterances were used to train monophone acoustic phone models and align them to the speech signal using a train-test-evaluate-retrain bootstrapping process (note that the segmentation is based purely on acoustics, and does not take into account factors such as phonotactic information). The 3-state models were trained using a 10 ms frame shift, resulting in a minimum duration of 30 ms per phone. Using a lexicon of canonical English pronunciations taken from the CELEX database (Baayen et al. 1996), which includes one or more pronunciations per word, the system chose the best pronunciation based on the speech signal for each word in the orthographic transcription, resulting in a phone-level transcription of the utterances. For each instance of *the*, the system chose one of the three variants listed in the dictionary: /ðə/, /ðʌ/, or /ði/. Since the difference in vowel quality between /ðə/ and /ðʌ/ is minimal, with the main distinction lying in suprasegmental properties not considered by the segmenter, we collapsed these two pronunciations into a single category, /ðə/.

We then extracted a total of 3180 tokens of *the* along with their phrasal context. We also manually verified a portion of the data to evaluate accuracy of the alignment and choice of vowel quality. Fifty tokens were randomly selected and aligned for vowel length, and these values were compared with the boundaries given by the automatic segmenter by taking the absolute values of the differences between the manual and automatic boundaries. The start points of the vowel had an average discrepancy of 5 ms (standard deviation = 4.06), while the end points had a greater discrepancy of 13 ms (standard deviation = 27.8). Excluding one extreme outlier of 188 ms difference, however, the discrepancy dropped to 9 ms for the end points. For vowel quality, 150 tokens were coded as /i/ or /ə/, then checked with the automatic transcription. The human annotation and the automatic transcription agreed on 148 out of the 150 tokens.

In order to examine the effect of frequency on the variability of *the* within fluent speech, tokens with immediate preceding or following disfluencies (i.e. silences, repetitions) were omitted from this study (we turn to these cases in Study 2). Some obvious hesitations are signaled only by an abnormally long vowel in the token of *the* itself: for example, a token of *the* with an 800 ms vowel should not be considered fluent speech. To find an objective cutoff point to exclude these ‘extra-long’ vowels from the data set, we examined the density plot of vowel durations, which showed a bimodal distribution with a valley at approximately 165 ms. Manual inspection of a random sample confirmed that tokens in the longer group sounded like hesitations, so we eliminated *the* tokens containing vowels longer than 165 ms. Table 1 lists the number of remaining tokens by speaker and language. Speakers with the same number were in the same conversation; the number of tokens from Speaker A is consistently higher because their task was to describe the picture to Speaker B, resulting in more total speaking time for Speaker A than for Speaker B.

Table 1: Number of fluent tokens of ‘the’ by speaker and language

Norwegian				**Czech**			
Speaker	*Tokens*	*Speaker*	*Tokens*	*Speaker*	*Tokens*	*Speaker*	*Tokens*
1a	164	1b	57	6a	151	6b	70
2a	27	2b	28	7a	183	7b	40
3a	170	3b	140	8a	108	8b	64
4a	198	4b	18	9a	283	9b	43
5a	301	5b	130	10a	56	10b	16
Total			**1233**				**1014**

Vowel duration (in ms) was measured from the phone-level transcription. The analyses below are based on the log vowel durations (we use the term ‘vowel duration' for simplicity). Vowel quality was also recorded based on the output of the automatic segmenter, resulting in 2100 tokens of /ə/ and 147 of /i/. Because all onset consonants were coded as /ð/ by the automatic segmenter, consonant quality was manually coded by the first author as either /ð/, another consonant, or ‘reduced’ (if there is no distinct consonant perceptible in the signal). To check for reliability, a subset of 150 of the consonants was also coded by another native-English-speaking phonetician: the two transcriptions showed over 85% inter-transcriber agreement.

Since the object of investigation is the word *the*, which in fluent speech obligatorily forms a prosodic, syntactic, and semantic unit with the following word, the frequency of this following word was considered a reasonable estimation of the frequency of the constituent including *the*. We used the log of the lemma frequency of the following word provided in the CELEX database (Baayen et al. 1996)[1].

Speech rate, which was used as a control factor in all of the models, was calculated by taking the duration of each of the four phones preceding and following the target *the* relative to the average duration for that phone over the entire phone-level transcription, then taking the mean of these eight values[2]. If there were fewer than four preceding phones, only the following phones were considered, and *vice versa*. Our evaluation of speech rate entails that negative values represent shorter phone duration and thus faster speech rates.

2.2. *Results*

We used mixed effects regression models to examine the effects of following word frequency on vowel duration, vowel quality, and consonant quality. In order to control for speaker- and word-specific effects, we included speaker and following word as random factors. The models also controlled for speaker gender (which did not reach significance for any of the models and was thus omitted from the models in the analyses below), native language (Czech or Norwegian), speech rate, and whether the following segment was a consonant or vowel[3]. This last factor derives from the fact that non-native speakers are often explicitly taught an allophonic alternation in which /ði/ occurs before a vowel and /ðə/ before a consonant (Fox Tree and Clark 1997, Keating et al. 1994), although there is considerable variation in how it is used in native speakers of English, which is influenced by other linguistic factors such as the stress pattern of the following word (Raymond et al. 2002, 2009) or social factors such as age (Keating et al. 1994). Additional control factors used specifically in the analysis of consonant quality will be discussed in the corresponding sections below.

2.2.1. *Vowel duration.* Vowel duration showed significant effects of language, speech rate, vowel quality, and following word frequency (Table 2). Unsurprisingly, vowels are longer in the context of slower speech rates, just as for native English speakers (Bell et al. 2003), and, holding all else equal, /i/s are 41 ms longer than /ə/s. Vowel durations were significantly longer in Czech than in Norwegian speakers (all else being equal, Czech vowels were on average 11 ms longer). There were also significant interactions of language with speech rate and following word frequency, prompting us to look at the two languages separately. For Czech speakers, all three factors were significant (speech rate: $\beta = 0.03$, $t = 4.44$, $p < .0001$; vowel quality: $\beta = 0.49$, $t = 9.95$, $p < .0001$; following frequency: $\beta = -0.02$, $t = -2.19$, $p < .05$), showing that vowels are shorter before a more frequent following word. For Norwegian speakers, on the other hand, the effect of following word frequency did not reach significance, but the effects of speech rate and vowel quality went in the same direction as those of the Czech speakers (speech rate: $\beta = 0.05$, $t = 7.70$, $p < .0001$; $\beta = 0.20$, $t = 4.31$, $p < .0001$; following frequency: $p > .05$).

Table 2: Significant factors affecting vowel duration in fluent speech: regression coefficients and *t*- and *p*-values

Factor	*β*	*T*	*p* <
vowel quality	0.74	13.55	.0001
following phone (C/V)	-0.49	-7.24	.0001
Language	-0.43	-4.70	.0001
speech rate	0.03	3.80	.01
following frequency	-0.02	-3.11	.01
language * vowel quality	-0.24	-3.78	.01
language * speechrate	0.03	3.28	.01
language * frequency	0.03	3.13	.001

The fact that the effect of frequency only held for one group of speakers suggests that the effect might be more complex than expected, or that the frequency measure used did not well reflect the words' frequencies for Norwegian speakers. Nevertheless, the results support the findings in Baker et al. (2011) that non-native speakers may show sensitivity to probabilistic effects in the same way as native speakers, in that function words in more frequent word pairs tend to be shorter.

2.2.2. *Vowel quality*. Results for vowel quality are shown in Table 3. We used a logistic regression model including following word frequency as well as the control factors of language, gender, speech rate, and following segment class. We found that only the factor of following segment class had a significant effect. In particular, in the case of a following consonant, speakers used /ə/ 96% of the time, while in the case of a following vowel, /ə/ was used only 35% of the time[4]. No other significant effects were found.

Table 3: Significant factors affecting vowel quality in fluent speech: regression coefficients and *z*- and *p*-values

Factor	*β*	*z*	*p* <
following phone (C/V)	6.04	12.85	.0001

2.2.3. *Consonant quality*. Several speakers showed little or no variation in consonant quality, in that all or nearly all of their *the* tokens contained the canonical /ð/. In order to model variation, speakers for whom more than 95 percent of tokens contained [ð] were excluded from the analysis: we removed three speakers who had zero, one, and three non-[/ð/] tokens (one Czech, two Norwegians). The tokens that did not contain /ð/ were coded as either having different consonant (684 tokens) or having a fully reduced consonant (no consonant perceptible in the signal, 62 tokens).

We again investigated the potential effect of following word frequency, while controlling for speech rate, vowel duration, vowel quality, language, and gender. In addition, the articulation of the preceding phone might be expected to affect the quality of the consonant: in particular, we expect that the consonant in *the* assimilates to the place of articulation of the preceding consonant, in native as well as non-native speech (e.g. Zhao et al. 2010). Since all of the substitutions for /ð/ observed in the data were alveolar consonants ([d], [t], [ɾ], [z], [n]), these substitutions may result from assimilation if the preceding consonant is alveolar. We investigated the role of assimilation by incorporating a factor with two levels, 'alveolar' or 'not alveolar' (which included preceding silence, making up 15% of the non-alveolar tokens). We initially excluded completely reduced consonants (those not visible in the signal) from the analysis, as they could not be considered either assimilated or unassimilated.

Logistic regression revealed effects of preceding phone class, vowel quality, and following word frequency. Results are shown in Table 4. Place of articulation of the preceding phone had the expected effect: the consonant is more likely to be one of the alveolar variants (and thus less likely to be [ð]) when the preceding phone is an alveolar consonant. When the preceding phone is not an alveolar consonant, [ð] occurs 78% of the time, versus 73% of the time when the preceding phone is alveolar. Vowel quality is also a significant predictor of consonant quality: the consonant is more likely to be the fricative [ð] when the vowel is the full vowel /i/

(81% [ð]) rather than /ə/ (76% [ð]). The following phone class had an effect as well: [ð] was more common before a vowel (88% [ð]) than before a consonant (76% [ð]). Finally, the consonant is less likely to be [ð] when followed by more frequent words.

Table 4: Significant factors affecting consonant quality in fluent speech: regression coefficients and *z*- and *p*-values

Factor	*B*	*z*	*p* <
vowel quality	1.75	4.26	.0001
following coronal	-0.37	-2.58	.001
following phone (C/V)	-0.36	-2.08	.05
following frequency	-0.07	-2.06	.05

When the completely reduced tokens were put back into the data set, preceding phone class had no effect on whether the consonant was [ð] or not-[ð]. This is unsurprising, because there is no reason to believe that a neighboring alveolar would condition deletion of a segment, so the addition of the completely reduced consonants obscures the effect of alveolar assimilation shown above. However, the two other factors of vowel quality and following word frequency were again significant, as was vowel duration, in that in the case of longer vowels, the consonant was more likely to be /ð/ (duration: $\beta = 0.42$, $z = 3.23$, $p <.01$; vowel quality: $\beta = 0.58$, $z = 2.12$, $p < .05$; following word frequency: $\beta = -0.08$, $z = -2.04$, $p < .05$).

In sum, /ð/ is more likely to be substituted by another consonant, or reduced, before more frequent words. Substitution and reduction of /ð/ are also conditioned by more reduced vowels, while reduction is additionally conditioned by shorter vowels, suggesting that these substitutions are more likely to occur in contexts conducive to articulatory reduction.

3. Study 2: Effects of disfluencies on non-native English *the*

The second study investigates the effect of disfluencies on non-native speech. If effects for non-native speakers mirror those of native speakers, we expect that *the* will be longer and more likely to contain the full vowel /i/ in the context of a disfluency (Fox Tree and Clark 1997, Bell et al. 2003). Furthermore, since we found that the consonant of *the* is less likely to reach the target [ð] in contexts that favor reduction, we might expect that the consonant is more likely to be pronounced [ð] in the context of disfluencies.

The types of disfluencies and their distribution used in this study are given in Table 5. We focus on disfluencies following the target word *the* because many preceding disfluencies are difficult to determine automatically: for example, *the* followed by a silence can be assumed to be a hesitation because of its close syntactic and prosodic relationship with the following word, but a silence preceding *the* could simply be a pause at a prosodic boundary. Since Bell et al. (2003) found that a model including following disfluency as a factor predicts vowel duration better than a model using preceding disfluency, Furthermore, findings from Bell et al. (2003) suggest that following disfluencies have a greater effect on a word’s form than preceding disfluencies, so examining following disfluencies should be the optimal way to reveal effects, if they exist..

Table 5: Distribution of following disfluency types in this study

Disfluency type	*Tokens*	*Example*
Silence (> 100 ms)	350	‘from **the [*silence*]** tip’
Repetition of *the*	109	‘the girl and **the the** dog’
Filled pause	75	‘rest of **the um** rest’
Repair	68	‘on **the of** the page’
Partial word	55	‘**the sh-** oh yeah the shape’
Long vowel	420	‘**th[e:]** right side’

3.1. *Method*

We added the previously omitted tokens with disfluencies and long vowels to the data subset from Study 1, which resulted in 3180 tokens. We again used regression analyses to model variation in vowel duration, vowel quality, and consonant quality. The predictor of interest in each case was the presence versus absence of a following disfluency, and we included as control factors the predictors that were found to have effects on each variable in Study 1. The following phone was counted as the sound following the target *the*, regardless of the type of disfluency; for example, the following sound for the filled pause example in Table 5 is considered the /ʌ/ of ‘um,’ while the following sound for repairs was the first sound of the incorrect word (‘of’ in the example above), as opposed to the first sound of the word eventually produced (‘page’ in the example above). Frequency was not included as a predictor because we did not have access to frequency counts for most following hesitations (e.g. silence, partial words, filled pauses).

3.2. *Results*

3.2.1. *Vowel duration.* Results for vowel duration are shown in Table 6. For the analysis of vowel duration, we did not consider tokens with an ‘extra-long’ vowel (as defined above) to be disfluencies unless they contained an additional type of disfluency as well. We omitted these tokens in order to see if disfluencies cued by something other than vowel length alone have an effect on vowel duration. There was a very large effect of disfluency even without counting extra-long vowels as disfluent: vowels are on average 120 ms longer in the presence of a following disfluency. An interaction between language and disfluency showed that the effect of disfluency is larger for Norwegian than for Czech speakers, so that even though Norwegian vowels were shorter overall (96 ms as opposed to 101 ms for Czech speakers), in the case of a disfluency the vowel durations are not significantly different between the two language groups. This might be expected if hesitations are symptoms of planning problems: while Norwegian vowels are shorter in general, the planning process takes equally long in both languages, resulting in similar vowel duration when there are planning problems. Otherwise, we found the same effects as in Study 1 for speech rate (vowel duration is longer in a slower speech rate) and vowel quality (/i/s are longer than /ə/s).

Table 6: Significant factors affecting vowel duration: regression coefficients and *t*- and *p*-values

Factor	*β*	*t*	*p* <
following disfluency	1.11	31.20	.0001
speechrate	0.04	7.74	.0001
vowel quality	0.32	9.11	.0001

language	-0.23	-3.89	.001
disfluency * language	0.21	4.59	.0001

3.2.1. *Vowel quality.* Based on findings for native English (Bell et al. 2003, Fox Tree and Clark 1997), we might expect non-native tokens of *the* which precede disfluencies to be more likely to contain the full vowel /i/. Adding the factor of disfluency to the model from Study 1 resulted in no main effect of disfluency, but there was a significant interaction between disfluency and following phone class showing that the effect of disfluency goes in different directions for the two following phone classes (β = -3.79, z = -4.32, p < .0001). The data was then split into fluent and disfluent subsets. The fluent data (analyzed above) showed an effect of following phone class (β = 6.04, z = 12.85, p < .0001), with /i/ being more likely before a vowel than a consonant. For the disfluent data, /i/ was also significantly more likely before a vowel (β = 1.57, z = 3.55, p < .001). However, the “following vowel” of the disfluent data includes filled pauses such as “um” and “uh,” which might be expected to behave differently. Therefore, we split the condition of following phone class into four levels: consonant, vowel, filled pause, and silence. Regression analysis showed that there was a significant effect of following phone class: filled pauses were more likely to be /i/ than the other levels (β = 2.55, z = 5.01, p < .0001) (see Figure 1).

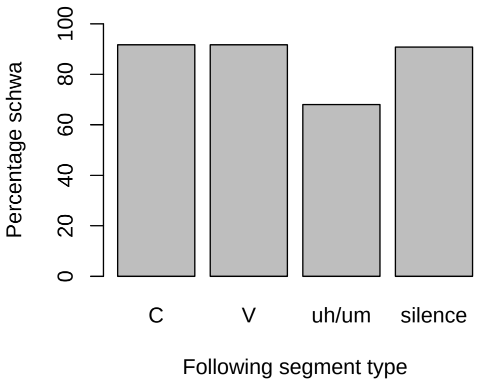

Figure 1: Percentage of tokens with vowel quality /ə/ in disfluent contexts, sorted by following segment type.

These results suggest that it is not the case that these non-native speakers generally use /ði/ in the context of disfluencies to signal planning problems, contrary to findings from native speakers (Fox Tree and Clark 1997): in our data set /ə/, not /i/, was produced the majority of the time in disfluent contexts (regardless of following phone). We think that the pattern in Figure 1 can be accounted for with the assumption that these non-native speakers generally follow the allophonic distinction: /i/ signals above all a following vowel, while /ə/ signals a following consonant. In the case of a disfluency, the speaker may not yet know whether the following phone is a vowel or a consonant. If this is the case, the bias towards /ə/ shown in our results is to be expected, taking into account that there are more words starting with consonants than vowels (see also Biber et al. 1999:1059 for a similar phenomenon with the indefinite article in repairs made by native English speakers). The fact that speakers used /i/ more before filled pauses may be due to the fact that they are aware a (vowel-initial) filled pause will follow ‘the.

3.2.1. *Consonant quality*. Results for consonant quality are shown in Table 7. The analysis of the effect of following disfluency on consonant quality shows that [ð] is more likely in the context of a following disfluency: 77% of consonant tokens within fluent speech are realized as [ð], while 81% are [ð] in the context of a following disfluency. As in Study 1, the consonant in *the* is also more likely to be a fricative when it is followed by the full vowel [i] rather than [ə]. Also as in Study 1, [ð] is also more likely to be substituted with a different consonant when preceded by an alveolar consonant. However, a significant interaction between preceding alveolar and disfluency shows that in the context of a following disfluency, the effect of preceding alveolar consonant disappears (β = -0.23, z = -1.91, p > .05). These patterns support our hypothesis that [ð] is more likely in the context of disfluencies.

Table 7: Significant factors affecting consonant quality: regression coefficients and *z*- and *p*-values

Factor	β	z	p <
vowel quality	0.70	3.63	.001
preceding coronal	-0.34	-2.38	.05
following disfluency	0.33	2.00	.05
preceding coronal * disfluency	0.71	2.10	.05

4. Discussion and Conclusion

This work has investigated whether non-native speakers show native-like reduction patterns and effects of disfluency in spontaneous speech. In particular, we have studied the effects of following word frequency and following disfluencies on the variability of vowel duration, vowel quality, and consonant quality in tokens of *the* in conversational English spoken by native speakers of Czech and Norwegian.

Our results showing that *the* is shorter preceding more frequent words suggest that speakers are sensitive to lexical frequency effects in conversational tasks, even in their second language. This supports similar findings on the effects of frequency in non-native read speech (Baker et al. 2011), extending them to conversational speech. Furthermore, we found that speakers also show native-like effects of longer vowels in the context of disfluencies. This finding suggests that lexical frequency and disfluencies may have similar effects on duration cross-linguistically, which is expected if these effects result from planning processes; if this is the case, then non-native speakers can use the same processing and production mechanisms as in their native language, and even less proficient speakers might be expected to show these effects.

The relatively proficient non-native speakers in our study also showed a native-like allophonic distinction for vowel quality, in that /i/ occurs much more often before vowels, while /ə/ occurs more often in other contexts. However, they diverged from previous findings on native speakers in that /ə/, rather than /i/, occurs much more frequently in the context of disfluencies. Fox Tree and Clark (1997) propose that the use of /i/ in native English is a ‘signal' of an upcoming disfluency. They argue that this signal must be planned in advance, as opposed to simply being the symptom of an online production problem. If it is true that the /ði/ *vs.* /ðə/

distinction has this higher-level pragmatic function, then it is clearly a language-specific one. It is unlikely that such a subtle language-specific distinction would be a salient cue for learners of English, and therefore even highly proficient non-native speakers would be expected to have trouble learning the distinction, which is in line with our findings.

The consonant [ð] was more likely to be substituted by a different consonant in more frequent word pairs, while in contexts that do not favor reduction (i.e. before disfluencies), the consonant was more likely to be pronounced as the target [ð], as opposed to a substitution by another consonant ([d], [t], [n], [ɾ], or [z]) or complete reduction of the consonant. This result suggests that at least for non-native speakers, reduction of [ð] can be realized by a change in both place and manner of articulation. A direct comparison with native speech patterns is not possible in this study, because previous work has not considered the distribution of [ð] and its substitutes in terms of lexical frequency and disfluencies. However, when comparing our results to those of (Zhao 2007), who found that stop-like substitutions for [ð] were more frequent in read than in spontaneous speech, it appears that the patterns of variability in the production of non-native English [ð] are different than those of native speakers: Zhao's (2007) findings suggest that another consonant is substituted for [ð] more often in more casual (spontaneous) speech than in more formal (read) speech, suggesting that it might not be articulatory reduction, whereas the patterns of [ð]-substitution found in this corpus seem to mirror contexts in which articulatory reduction is expected to occur.

Interestingly, although we studied speakers from two distinct language backgrounds, we found few language-specific differences. The differences we did find were all related to vowel duration: Norwegian speakers have shorter vowel durations in general, and effects of slower speech rate and following disfluency, both of which make vowels longer for all speakers, were larger for Norwegian than for Czech speakers. The effect of following disfluency is expected if disfluencies are caused by planning problems: although vowel durations are usually shorter for Norwegian speakers than for Czech speakers, the effect of planning problems is probably equivalent across the two languages, and the problems should therefore take equally long to resolve, resulting in roughly equivalent vowel duration. Furthermore, the effect of following word frequency is not significant for vowel duration in Norwegian speakers. There may be other language-specific effects, but due to the heterogeneity characteristic of non-native speech (Baker et al. 2011), such effects may be obscured, and larger or more controlled data sets may be necessary to reveal them.

Although our results are preliminary and must be extended to a larger set of words and language backgrounds in order to be truly generalizable, this study has identified several similarities and differences between native and non-native sensitivity to frequency factors and effects of disfluencies in English spontaneous conversations. The non-native speakers in our corpus showed native-like durational patterns, suggesting that the impact of lexical frequency and disfluencies on function words may be similar for native and non-native speakers of English. On the other hand, native-like use of vowel quality to signal a higher-level communicative function was not used by the non-native speakers. The similarities between native and non-native speakers follow if effects of predictability and disfluency result from a non-language-specific planning process, which should have the same effect regardless of the language being spoken, whereas the differences between native and non-native speakers with respect to vowel quality in disfluent contexts suggest that the native English distribution of vowels in *the* in these contexts does not result from an automatic process common to speakers of all languages, but is learned. In sum, we have shown that corpus study of non-native pronunciation variation can provide not only observations about the distribution of this variation, but also insights into reduction processes in foreign and native language speakers.

Acknowledgements

Thanks to Louis ten Bosch for his help with HTK and to Dan Brenner for his help with verification of transcriptions. Thanks also to two anonymous reviewers for valuable advice and suggestions. This work was partially funded by the Marie Curie Research Training Network *Sound to Sense* and a European Young Investigator Award to the second author.

Bionotes

Jessamyn Schertz is a Ph.D. student in the Department of Linguistics at the University of Arizona. Her research focuses on phonetic cue weighting in perception and production. The present work was completed during her participation in the Marie Curie Research Training Network *Sound to Sense*.

Mirjam Ernestus is a full professor of Psycholinguistics at Radboud University Nijmegen. She is a specialist in the production and comprehension of acoustic reduction. Her current research projects focus on how non-native listeners understand acoustic reduction.

Notes

[1] Based on a suggestion from an anonymous reviewer, we also examined the effect of word bigram frequency (instead of following word frequency), based on the Contemporary Corpus of American English (Davies 2008). This did not change the outcome of the analyses.

[2] As pointed out by a reviewer, the duration of the following phones may also be influenced by following word frequency, potentially reducing the apparent effect of frequency. We therefore redid the analyses by normalizing the duration only on the basis of the preceding (but not following) phones; however, this did not change the outcome of the analyses.

[3] Since the following segment class was expected to correlate highly with vowel quality, the residuals of the correlation between vowel quality and following segment class, as opposed to the raw data, were used for the vowel duration and consonant quality analyses.

[4] Many non-native speakers are explicitly taught this alternation (see Keating et al. 2004).

References

Aylett, Matthew and Alice Turk. 2004. The smooth redundancy hypothesis: a functional explanation for the relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47. 31-56.

Baayen, R. Harald, Richard Piepenbrock, and L. Gulikers. *CELEX2*. Linguistic Data Consortium, Philadelphia, 1996.

Baker, Rachel E., Melissa Baese-Berk, Laurent Bonnasse-Gahot, Midam Kim, Kristin J. van Engen, and Ann R. Bradlow. 2011. Word durations in non-native English. *Journal of Phonetics* 39(1). 1-17.

Bell, Alan, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Daniel Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 92-111.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English. *Journal of the Acoustical Society of America* 113. 1001-1024.

Biber, Douglas, Stig Johannson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Cao, Ying Alisa. 2002. Analysis of acoustic cues for identifying the consonant /ð/ in continuous speech. Master's thesis, MIT.

Ernestus, Mirjam and Natasha Warner. 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics* 39. 253-260.

Fox Tree, Jean E. and Herbert H. Clark. 1997. Pronouncing "the" as "thee" to signal problems in speaking. *Cognition* 62(2). 151-167.

Jescheniak, Jörg D. and Willem J.M. Levelt. 1994. Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(4). 824-843.

Johnson, Keith. 2004. Massive reduction in conversational American English. In K. Yoneyama and K. Maekawa (eds.), *Spontaneous speech: data and analysis*. The International Institute for Japanese Language, Tokyo. 29-54.

Manuel, Sharon. 1995. Speakers nasalize /ð/ after /n/, but listeners still hear /ð/. *Journal of Phonetics* 23(4). 453-476.

Pluymaekers, Mark, Mirjam Ernestus, and R. Harald Baayen. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62(2-4). 146-159.

Spilková, Helena, Daniel Brenner, Anton Öttl, Pavel Vondricka, and Wim van Dommelen. 2010. The Kachna L1/L2 Picture Replication Corpus. *International Conference on Language Resources and Evaluation*, Valletta, Malta.

Young, Steve, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2002. *The HTK Book*. Cambridge University Engineering Department, Cambridge.

Zhao, Sherry. 2007. The stop-like modification of /ð/: a case study in the analysis and handling of speech variation. PhD thesis, MIT.

Zhao, Sherry. 2010. Stop-like modification of the dental fricative /ð/: an acoustic analysis. *Journal of the Acoustical Society of America* 128. 2009-2020.