Predicting Acoustically Reduced words in Spontaneous speech: The role of Semantic/Syntactic and Acoustic cues in context

Marco van de Ven[1,2], Mirjam Ernestus[1,2],
and Robert Schreuder[3]

Corresponding author:

Marco van de Ven

Radboud University Nijmegen

P.O. Box 9103

6500 HD Nijmegen

The Netherlands

Telephone: +31-24-3615752

E-mail: Marco.vandeVen@mpi.nl

[1] Centre for Language Studies, Radboud University Nijmegen, The Netherlands
[2] Max Planck Institute for Psycholinguistics, The Netherlands
[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands

Abstract

In spontaneous speech, words may be realised shorter than in formal speech (e.g., English *yesterday* may be pronounced like [jɛʃeɪ]). Previous research has shown that context is required to understand highly reduced pronunciation variants. We investigated the extent to which listeners can predict low predictability reduced words on the basis of the semantic/syntactic and acoustic cues in the context. In four experiments, participants were presented with either the preceding context or the preceding and following context of reduced words, and either heard these fragments of conversational speech, or read their orthographic transcriptions. Participants were asked to predict this missing reduced word on the basis of the context alone, choosing from four plausible options. Participants made use of acoustic cues in the context**,** although casual speech typically has a high speech rate, and acoustic cues are much more unclear than in careful speech. Moreover, they relied on semantic/syntactic cues. Whenever there was a conflict between acoustic and semantic/syntactic contextual cues, measured as the word's probability given the surrounding words, listeners relied more heavily on acoustic cues. Further, context appeared generally insufficient to predict the reduced words, underpinning the significance of the acoustic characteristics of the reduced words themselves.

Index terms: speech reduction, word recognition, semantic/syntactic predictability, acoustic context, casual speech.

Acknowledgements

Mirjam Ernestus was supported by a European Young Investigator Award from the European Science Foundation. We would also like to thank Natasha Warner for helpful comments on the use of square wave signals in perception experiments and Francisco Torreira for providing a script to create a square wave signal for our experiments.

1. Introduction

In casual speech, words may be pronounced much shorter in duration and in terms of the number of clear and distinct segments than in formal speech (for an overview, see Ernestus and Warner, 2011). For example, in casual English the adjectives *ordinary* and *hilarious* with the citation forms [ɑrdənɛrɪ] and [hɪlɛriəs] can be pronounced as [ɑnrɪ] and [hlɛres], respectively (Johnson, 2004). The absence of single or multiple segments is very common in casual speech. In fact, Johnson (2004) reported that, in English, complete syllables are absent in 6% of the tokens on average. The present study investigated how listeners use context in order to comprehend reduced variants.

Previous research has shown that listeners need context to recognise reduced pronunciation variants, and not just the sounds that pertain to the reduced variants themselves (see also Bard, Shillcock, and Altman, 1988 for similar results on spontaneous but not necessarily reduced speech). Ernestus, Baayen, and Schreuder (2002; and see also Janse and Ernestus, 2011), for example, presented Dutch listeners with sound fragments extracted from the Ernestus Corpus of Spontaneous Dutch (Ernestus, 2000), including pronunciation variants of a low, medium, or high degree of reduction. These variants were either presented in isolation (i.e. only that part of the speech signal for which most acoustic cues are related to the target word, not to the context), in phonological context (the neighbouring vowels and any intervening consonants), or in sentential context, and the participants were asked to orthographically transcribe the speech fragments presented. Their results showed that listeners have difficulty identifying highly reduced pronunciation variants in isolation (ca. 50% correct). More importantly, although their performance increased significantly when highly reduced variants were presented within their phonological context, identification problems remained (ca. 70% correct). Only when presented with the full sentence context, listeners were able to identify highly reduced variants successfully (more than 90% correct). For the variants with a medium degree of reduction, listeners needed only the phonological context for correct identification, while the variants with a low degree of reduction were identified correctly in all three context conditions. This study then suggests that, although phonological contextual information is beneficial, a larger context is required for the

identification of highly reduced variants, while this is not the case for less reduced variants. This finding may be explained by the fact that less reduced pronunciation variants are more similar to the variants that occur in single-word utterances than more heavily reduced variants. The question remains, however, how listeners use context to recognise highly reduced pronunciation variants.

In the present study, we focused on the roles of acoustic context and semantic/syntactic context during the processing of reduced pronunciation variants. By acoustic context we mean the acoustic signal that is left after splicing out that contiguous part of the signal that contains cues that mostly pertain to the segments of the target word. Obviously, it is impossible to splice out a word completely from the speech signal, since cues are typically also present in surrounding words (i.e. contrary to the *Absolute Slicing Hypothesis*; Goldsmith, 1976). The surrounding words may contain cues resulting from coarticulation of adjacent sounds. For instance, the realisation of the nasal in the word combination *garden bench* may contain information about the place of articulation of the next word's initial consonant. Similarly, due to vowel-to-vowel coarticulation, the final vowel in a word may provide information about the vowel in a following word (Martin and Bunnell, 1982). The acoustic signal may even contain acoustic cues for segments that are several syllables away (e.g. Kelly and Local, 1986). West (1999, 2000), for instance, has shown that there are differences in *F*3 and in the position of the tongue, both for /r/ and /l/, in vowels that are not adjacent to these liquids, and that listeners are sensitive to this fine phonetic detail. In addition, the duration of a vowel may help listeners to interpret the first vowel of a following word as phonologically long or short (Nooteboom, 1972). Finally, prosody (the alternating pattern of stressed and unstressed syllables, as well as intonation) may form cues to the prosodic characteristics of an upcoming word. In the present study, we investigated to what extent this variety of acoustic cues in the context helps listeners to predict reduced pronunciation variants.

Some researchers predict that listeners pay more attention to any type of acoustic information in conversational speech than in laboratory speech, since conversational speech is fast, often reduced and accompanied with background noises, and listeners would use all cues available

under such adverse listening conditions (Hawkins and Smith, 2001). In line with this hypothesis, Heinrich, Flory and Hawkins (2010) found that listeners are more sensitive to the acoustic traces of /r/ in syllables preceding /r/ (r-resonance, Kelly and Local, 1986), especially in spontaneous speech. However, acoustic cues may also be too subtle to help listeners predict words in conversational speech. These cues may be difficult to notice and they may be difficult to process also due to the high speed and variability of this speech register.

In addition to the role of the acoustic cues, we investigated the role of the semantic/syntactic cues in the context during the processing of highly reduced pronunciation variants. By semantic/syntactic context we mean all information that can be extracted from the orthographic transcription of the context. Since hardly any research has investigated exactly how semantic/syntactic cues in the context are used to recognise reduced pronunciation variants, we need to consult the literature on the comprehension of carefully pronounced words to find out which semantic/syntactic cues might contribute to the recognition of these reduced variants.

To begin with, in a cross-modal priming study, Zwitserlood (1989) investigated at which stage in the comprehension process the influence of semantic context sets in. Dutch participants were presented with prime words that have a relatively late uniqueness point (e.g. Dutch /kapit/ is consistent with both /kapi'tɛin/ "captain" and /kapi'tal/ "capital", so the uniqueness point is after the /t/). These words were embedded in a neutral context (e.g. *They mourned the loss of their captain.*) or a biasing context (e.g. *With dampened spirits the men stood around the grave. They mourned the loss of their captain.*). Participants were asked to make a lexical decision for visual probes that were either semantically related (for the example above: *ship*) or unrelated (for the example above: *money*) to the prime words in the auditorily presented sentences. In the biasing context condition, there was already significant priming for the semantically congruent word just before the uniqueness point of the auditory prime. These results indicate that semantic information in the preceding context can enhance word recognition already before a word becomes unique.

Similar conclusions were drawn by van den Brink, Brown, and Hagoort (2001, 2006), who

recorded event-related potentials (ERPs) while participants were presented with spoken sentences. The sentences ended either in semantically plausible (e.g. *The painter colored the details with a small brush.*) or implausible words (e.g. *The painter colored the details with a small pension.*). Contextually incongruent words yielded larger ERPs than contextually congruent words. More importantly, the onset of this N400 effect occurs prior to the word's uniqueness point, which indicates that the N400 peak is not simply due to semantic integration difficulty. Rather, listeners unconsciously formulate predictions about the upcoming words, and processing is inhibited if these predictions are incorrect.

In addition to the question of how the semantic/syntactic context is used, there has been research investigating what types of semantic/syntactic information in the context listeners can use to facilitate the word recognition process. Most language models that are used to model semantic/syntactic context effects in word recognition are based on N-gram frequencies (corpus-based frequency counts), and more specifically word trigram frequencies. Despite their simplicity, N-gram language models turn out to be very powerful tools in natural language processing (Wiggers, 2008). In addition, they well predict human word recognition (e.g. Morton and Long, 1976; McDonald and Shillcock, 2003). For example, McDonald and Shillcock (2003) conducted an eye-tracking study in which participants read sentences containing target words with high or low transitional probabilities with their preceding words. Their results showed earlier and shorter fixations on target words with a high N-gram frequency with their preceding words compared to target words with a low N-gram frequency with their preceding words. This finding indicates that N-gram models also partially reflect the way listeners can use semantic/syntactic information from surrounding words to facilitate word recognition.

So far, only one study has explicitly investigated the role of semantic contextual information that is carried by reduced pronunciation variants. Van de Ven, Tucker, and Ernestus (2011) conducted a series of auditory lexical decision experiments, with implicit semantic priming, in which both the primes and targets could be reduced or unreduced (all combinations were investigated). The results indicated that after reduced primes, participants needed more processing time (compared to unreduced primes) before they could use the semantic information

of the prime to facilitate the recognition of the upcoming target. Because of this delay in priming, it is unclear whether listeners can actually use the various types of information in the context when they are presented with natural, conversational speech, and to what extent.

In the present study, we investigate how semantic/syntactic and acoustic cues in the context help language users predict the identity of acoustically reduced words (or fixed expressions, henceforth *target words*, for the sake of convenience), embedded in their natural, reduced contexts. In four main experiments, participants were only provided with the preceding (Experiments 1 and 2) or the preceding *and* following context (Experiments 3 and 4). Participants either read the orthographic transcriptions of the context (Experiments 1 and 3), or they heard the acoustic signal of this context (Experiments 2 and 4). We will investigate the contribution of semantic/syntactic cues in the context on the basis of the performance by the participants in Experiments 1 and 3, and by testing the effects of unigram and N-gram frequency in all four experiments. We will focus on the contribution of acoustic cues by comparing Experiments 1 and 2, and Experiments 3 and 4.

We address various specific research questions with respect to the contribution of context to the predictability of reduced pronunciation variants. First of all, we tested whether listeners can use the semantic/syntactic and/or the acoustic context at all to predict reduced pronunciation variants. As explained above, this is by no means obvious, given that many words are pronounced quickly and segments tend to be missing in casual speech, which obscures acoustic cues to upcoming words and delays semantic priming. Further, we investigate whether the context is more informative if it contains a larger number of words.

Second, we will investigate which characteristics of the acoustic signal may help listeners predict upcoming reduced words. We hypothesize that listeners are more sensitive to acoustic detail if speech rate is low, and listeners thus have sufficient time to process these cues. Furthermore, there may be a difference between word-final vowels and consonants in how well they predict the onset of the following word. Transitional cues may be more salient and more informative in vowels than in consonants and therefore listeners may find it easier to predict upcoming words

that are preceded by words ending in vowels.

Third, we will investigate what types of semantic/syntactic contextual information can be used by listeners to predict reduced pronunciation variants. We will investigate whether this semantic/syntactic information can be completely captured by N-gram probabilities, and whether the role of this semantic/syntactic information is influenced by whether participants also have access to the acoustic cues in the context.

Finally, the degree of reduction of the target word may influence the extent to which listeners can use the semantic/syntactic context to predict reduced pronunciation variants. Listener driven accounts of speech reduction propose that speakers reduce especially those words that are highly predictable for the listener (Boersma, 1998). Hence, we expect that listeners can make better use of context to predict relatively highly reduced words than to predict relatively mildly reduced words.

2. Materials in the experiments

The materials used in the experiments were taken from the Ernestus Corpus of Spontaneous Dutch (Ernestus, 2000), consisting of casual conversations between ten pairs of speakers recorded in a soundproof booth. Since high frequency words are more likely to be reduced (e.g. Zipf, 1935; Bybee, 2001), we chose sixteen highly frequent polysyllabic Dutch words or word combinations as target words: *alleen* "only", *allemaal* "all", *altijd* "always", *anders* "otherwise", *bepaalde* "certain", *bijvoorbeeld* "for instance", *eigenlijk* "actually", *gewoon* "usual", *helemaal* "totally", *in ieder geval* "in any case", *misschien* "perhaps", *namelijk* "namely", *natuurlijk* "of course", *op een gegeven moment* "at a certain moment", *over* "about", and *tenminste* "at least". All of these words are adjectives, adverbs, or adverbial phrases and can be left out of their sentences without rendering these semantically incoherent or ungrammatical.

We selected on average five tokens (from different speakers) for each target word (mean: 4.88 tokens per word, range: 1 to 8 tokens). These tokens had low trigram frequencies with their two

preceding words or preceding and following word (2.97 and 1.52 per million respectively in the Spoken Dutch Corpus, Oostdijk, 2002) compared to previous studies. We added 22 filler word tokens which differed from the target tokens only in that they represented different word types, which introduced more variation (and therefore a smaller predictability of the correct option) in the experiment. In the end, the experiment consisted of 78 target word tokens and 22 filler word tokens, produced by eleven speakers.

We extracted the target and filler tokens together with some part of their preceding and following contexts. The amount of context varied for each token, but comprised at least the prosodic phrase in which the token was embedded. On average, the preceding context consisted of 9.63 words (range: 2 to 22, and one token with 29 words), and the following context consisted of 5.05 words (range: 1 to 13 words). The extracted fragments did not contain any overlapping speech or loud background noises. An orthographic transcription of the experimental materials is provided in Appendix A.

The degree of reduction of the target and filler tokens varied from mildly reduced (e.g. [ɛiχlək] for *eigenlijk*, with the unreduced pronunciation [ɛiχələk]) to highly reduced (e.g. [ɛik] for *eigenlijk*). Although the participants thus did not hear or see the target words themselves, a predictor of their performance may be the degree of reduction of this word (i.e. words that are more reduced may be easier to predict; see the fourth research question formulated above). We quantified degree of reduction (henceforth *Reduction Degree in Production*) by subtracting the number of segments in the reduced form from the number of segments in the citation form, and dividing its outcome by the number of segments in the citation form. Degree of reduction in our materials varied from 0 to 0.73. If the degree of reduction was equal to 0, there was still some reduction present, for example in the form of consonant or vowel lenition (e.g. full vowels produced as schwas). Target and filler tokens were labelled as highly reduced if *Reduction Degree in Production* was higher than 0.4 (based on the approximate bimodal distribution observed in *Reduction Degree in Production*); otherwise they were labelled as mildly reduced. This resulted in 42 target word tokens classified as mildly reduced and 36 as highly reduced.

In order to verify the intelligibility of the target and filler tokens within their contexts, we carried out a control experiment. In this experiment, twenty native speakers of Dutch, none of whom participated in the main experiments, first heard the reduced word in its original sentence context (e.g. *Ik vertrouw altijd maar op mijn goede geluk* "I always rely on good luck."), and then heard a shorter version of this same fragment, consisting of the same token of the reduced word and its two preceding and following words (e.g. *Ik vertrouw altijd maar op* "I always rely on")[4]. The participants were asked to orthographically transcribe this fragment. They were tested individually in a soundproof booth, on a PC running E-prime 1.2 (Schneider, Eschman, and Zuccolotto, 2002). The materials were grouped in eleven blocks, with each block containing the materials of one of the eleven speakers. Each block was preceded by a familiarisation phase, in which participants were presented with a short (on average 19 second) speech fragment of the speaker to get used to the speaker's (voice) characteristics. Subsequently, participants were presented with two filler tokens, followed by the target tokens. The results showed that three target tokens were difficult to understand (less than 60% correct), and they were not included in the analyses of the subsequent experiments. With respect to the remaining target tokens, participants successfully identified the mildly reduced tokens in 98% (range: 92.5% to 100% correct) and the highly reduced tokens in 94.8% (range: 60% to 100% correct) of the trials. These tokens can thus be well identified in their contexts.

We conducted a second control experiment, in order to investigate whether listeners could understand the target tokens used in our experiments in isolation. Participants listened to the reduced target and filler tokens without their contexts, and were asked to orthographically transcribe the words. The basic experimental procedure was adopted from the previous control experiment. Participants successfully identified the mildly reduced target tokens in 74% (range: 20% to 100% correct) and the highly reduced target tokens in 43.4% (range: 0% to 100% correct) of the trials. These recognition scores are very similar to those obtained by Ernestus et al. (2002). Thus, our two control experiments show that our reduced target tokens are difficult to

[4] In a few cases, the two preceding or following words were inseparable from their neighbouring words because these words had been contracted. In such cases, the context contained one or two (preceding or following) additional words. Conversely, the following context in some cases consisted of only one word because that word was sentence-final.

recognise in isolation but generally easy to understand in context.

We investigated whether the proportion of listeners correctly identifying a target token in isolation correlates with *Reduction Degree in Production* as defined above. We conducted a t-test which showed that *Intelligibility in Isolation* (the percentage of correct identifications in the second control experiment) differs between tokens that were classified as mildly and tokens that were classified as highly reduced (mean for mildly reduced tokens: 68%, mean for highly reduced tokens: 27%, one-tailed t-test: $t(73.6) = 6.27$, $p < 0.0001$[5]), which indicates that the two measures reflect a similar type of reduction. We decided to use *Intelligibility in Isolation* as a measure of reduction degree in our analyses of the main experiments, for various reasons. First, *Intelligibility in Isolation* reflects degree of reduction from a perception perspective, whereas *Reduction Degree in Production* reflects degree of reduction from a production perspective, and this study focuses on perception. Second, it is unclear whether the various types of reduction observable in the signal are equally important in perception (e.g. the absence of /r/ versus the absence of /k/), and consequently whether they should be equally important for a measure of degree of reduction. Third, *Reduction Degree in Production* only reflects the relative number of absent segments and does not take lenition (e.g. the pronunciation of full vowels as schwas) into account, but we know that also lenition may affect speech comprehension (e.g. Mitterer and Ernestus, 2006; Warner, Fountain, and Tucker, 2009). Fourth, it is particularly difficult to determine whether or not a segment has actually been realised (e.g. Ernestus and Baayen, 2011). As a consequence, a measure of reduction based on this information is somewhat unreliable.

We henceforth thus use *Intelligibility in Isolation* as a continuous measure of a word's degree of reduction although we know that a word's intelligibility is determined not only by its degree of reduction but also by other factors, for example its frequency of occurrence in natural speech contexts (e.g. Howes, 1954, 1957; Newbigging, 1961; Savin, 1963; Soloman and Postman, 1952). We believe that there is a strong relationship between a word's degree of reduction and its intelligibility in isolation. This assumption is supported by the correlation between *Intelligibility in Isolation* and *Reduction Degree in Production* reported above and by a global inspection of

[5] Correlation measures showed similar results.

the data from the second control experiment, which showed a large degree of variability between various tokens of the same word. For example, one token of *bijvoorbeeld* "for example" showed a success rate of 90%, whereas a different token of the same word showed a success rate of 27.5%.

After either hearing or reading the context (depending on the experiment), participants were presented with four semantically and syntactically plausible options, from which they had to select the target word which they believed to be located immediately after the preceding context in the original speech fragment. These four options were always provided orthographically. The four options included the correct answer, the word that followed the reduced target token in the original sentence, and two other options. An example trial is provided below.

Context:
Het geld is niet van jou en je staat _____
"The money is not yours and you are"

Options:
1. *altijd* "always" (correct, reduced word)
2. *rood* "in debt" (word following reduced word)
3. *bijvoorbeeld* "for instance"
4. *eigenlijk* "actually"

The original sentence for this example was *Het geld is niet van jou en je staat altijd rood*, and *altijd* "always" was thus the target word and *rood* "red" the following word. We included the following word as one of the four options because we initially wanted to verify whether semantic/syntactic and acoustic contextual information (e.g. prosody or formation transitions) can help listeners predict whether or not an additional word is present in a speech fragment.

The four options did not have identical word-initial sounds, making short-distance co-articulation cues potentially relevant in the auditory version of the experiment. Further, we made sure that, in

most cases, the correct answer was not the option with the highest probability in terms of lexical word frequency and N-gram frequency. The target word was the most frequent of the four options in only 5.25% of the trials and had the highest bi- or trigram frequency of the four options in only 14.47% of the trials. Hence, if participants just guess, all four options are equally likely. If participants are only sensitive to the lexical frequencies of the four options, they are expected to choose the word with the highest lexical frequency out of the four options and they therefore will perform 5.25% correct. If participants show some sensitivity to the preceding context and base their choice on the probability of the four options given the preceding word or the two preceding words (i.e. on bigram or trigram probability), they should also perform below chance level. Only if participants are sensitive to more semantic/syntactic information in the context than is captured by N-gram probability, they may perform above chance level. These frequencies were determined on the basis of the Spoken Dutch Corpus (Oostdijk, 2002).

By using a closed set of options we could more easily test which factors contributed to the predictability of words. Since we always provided several likely options given the context, we expect that we would have found similar effects if we had provided listeners with a larger set of words to choose from (which would then include less likely options), although these effects might have been smaller.

Most of the filler options also occurred as target words in the experiment. We can distinguish three types of words in the experiments, namely target words in target trials (these words occurred as options 17.81 times in the experiment on average), target words in filler trials (these occurred 1.76 times on average), and words that never served as targets in the experiment (each of these occurred 1.62 times on average). The order of the four options on the screen was randomised (between items and between participants). The order was manually corrected if it formed a semantically and syntactically plausible continuation of the sentence, which could happen for maximally 25 of the 78 trials. To illustrate with the example provided above, the order *bijvoorbeeld eigenlijk altijd rood* "for instance actually always in debt" would create a plausible continuation of the sentence, possibly leading to more *bijvoorbeeld* responses.

3. Experiment 1

In the first main experiment, participants read orthographic transcriptions of the contexts preceding the target words while these words themselves were missing. In addition, they saw four semantically/syntactically plausible options (as judged by the authors) for each target word and were asked to select the most likely one. The rationale behind this experiment was to establish how well participants could predict reduced words on the basis of only semantic/syntactic cues in the preceding context.

3.1 Participants

Twenty native speakers of Dutch from the pool of participants of the Max Planck Institute for Psycholinguistics (they were nearly all undergraduate students at the Radboud University Nijmegen) were paid to take part in the experiment. None of them reported any hearing loss.

3.2 Materials

Participants were provided with orthographic transcriptions of the context preceding the target word.

3.3 Procedure

The experiment consisted of eleven blocks, and each block contained the materials of one of the eleven speakers. The blocks and the trials within the blocks were randomised across participants. The experiment was self-paced, and was carried out in a soundproof booth. The experiment was programmed in and controlled by E-prime 1.2 (Schneider, et al., 2002).

For each trial, the preceding context was presented on the screen for five or eight seconds

(depending on the length of the sentence: Eight seconds if it consisted of more than sixteen words, otherwise five seconds), and then the four options appeared on the screen. Participants were asked to guess the following word by pressing one of the four buttons (labelled “1” to “4”) on a response box. The preceding context was then presented a second time so that the experiment was identical to the auditory experiment (see Experiment 2). The four options remained visible so that participants did not have to memorise the four options, which might otherwise interfere with their performance. Then, participants were asked to choose again.

3.4 Results and discussion

In all analyses presented in this study we investigated which variables favoured participants' selection of a given option by means of generalised linear mixed-effects models with the logit link function (see, e.g. Jaeger, 2008) and with random effects for *Participant*, *Target type* (the identity of the target word; e.g. namelijk or eigenlijk), and *Target token* (e.g. the third token of eigenlijk for a given participant in the experiment). We used a backwards stepwise selection procedure, in which predictors were removed if they did not attain significance at the 5% level. The fixed effect factors differed for each model, and will be mentioned for each model separately.

The descriptive statistics for Experiment 1 showed that participants, provided with four plausible options, selected the correct option in 33.27% of the trials, which is above chance (i.e. more than 25%, one-tailed t-test testing whether participants' performance was significantly higher than 25%; $t(19) = 5.58$, $p < 0.0001$). This was unexpected, since our target words were discourse markers and adverbs, which could be left out of the sentences. Apparently, listeners can use preceding semantic/syntactic information to also predict these types of words. Further, most target words had lower unigram and N-gram frequencies than at least one of the other three options on the screen. Thus, if participants had used the N-gram probabilities or the lexical frequencies of the four options, or if they had just guessed, we would have seen performance at or below chance level (0.25). The above chance performance therefore indicates that participants used semantic/syntactic information in the wider preceding context to predict the following word.

We analysed participants' selection of the correct answer versus the other options and included in the statistical model the fixed effects *Repetition* (whether the context was presented for the first or second time), position of the correct answer on the screen (henceforth *Position Correct*), and *Intelligibility in Isolation*. Since *Intelligibility in Isolation* was not distributed normally, we converted this variable into an ordinal variable with four levels, representing the four quartiles. The quartile ranges are presented in Table 1. This factor was included in all subsequent analyses

in this study (instead of the continuous variable).

Level	Range
Quartile 1	0% - 22.5%
Quartile 2	22.5% - 55%
Quartile 3	55% - 85%
Quartile 4	85% - 100%

Table 1: The quartile ranges for *Intelligibility in Isolation*.

We only found a main effect of *Position Correct*, which shows that the participants performed better if the correct answer was the second option and worse if the correct answer was the fourth option on the screen (F (3, 2996) = 32.04, $p < 0.0001$; 42.19% correct for option 2 and 22.44% correct for option 4). The following experiments also showed effects of *Position Correct*, suggesting that participants preferred Options 1 and 2 to Options 3 and 4. Since these position effects are not of primary interest for the present study, in the following sections their statistics will be reported in the tables, but they will not be discussed in the text. They are included in the statistical models to reduce the variance.

Importantly, the semantic/syntactic information was clearly insufficient for the participants to predict the target words without errors, as they chose incorrect options in no less than 66.71% of the trials. Given that listeners are sensitive to phonetic detail in laboratory speech, they may use acoustic cues in the preceding context to predict upcoming words in spontaneous speech as well. To investigate this possibility we conducted a second experiment, in which participants were auditorily presented with the preceding context of the target words. The properties of the reduced words may have influenced the realisation of neighbouring consonants/vowels or the prosody of the context. If participants use these acoustic cues, they should perform better in this experiment than the participants in Experiment 1.

4. Experiment 2

4.1 Participants

Twenty native speakers of Dutch from the pool of participants of the Max Planck Institute for Psycholinguistics were paid to take part in the experiment. None of the participants had taken part in any of the previous experiments, and none of them reported any hearing loss.

4.2 Materials

Experiment 2 is identical to Experiment 1, except that the preceding contexts were presented auditorily. Thus, listeners heard the preceding context of reduced target words, without the target words themselves (although there are acoustic cues in the context that belong to the target word, as mentioned previously). Despite the absence of clear word boundaries in conversational speech, we have tried to approximate these word boundaries (on the basis of the wave form and listening). Given that the informativeness of the context depends greatly on the placement of the word boundaries for the target words (i.e. the more acoustic information is spliced out, the less informative the context becomes), we have used strict guidelines for this procedure. Transcribers placed the boundaries of fricatives at the onset and offset of frication noise, the left boundary of plosives was placed at the onset of the closure, whereas the right boundary was placed directly after the burst. The location of the boundaries of nasals and liquids was determined by listening and by looking for sudden changes in the waveform. The word boundaries for our stimuli were cross-checked by two independent labellers, and verified by the first author. We provide examples of the segmentation procedure (solid lines indicate word boundaries) in Figure 1.

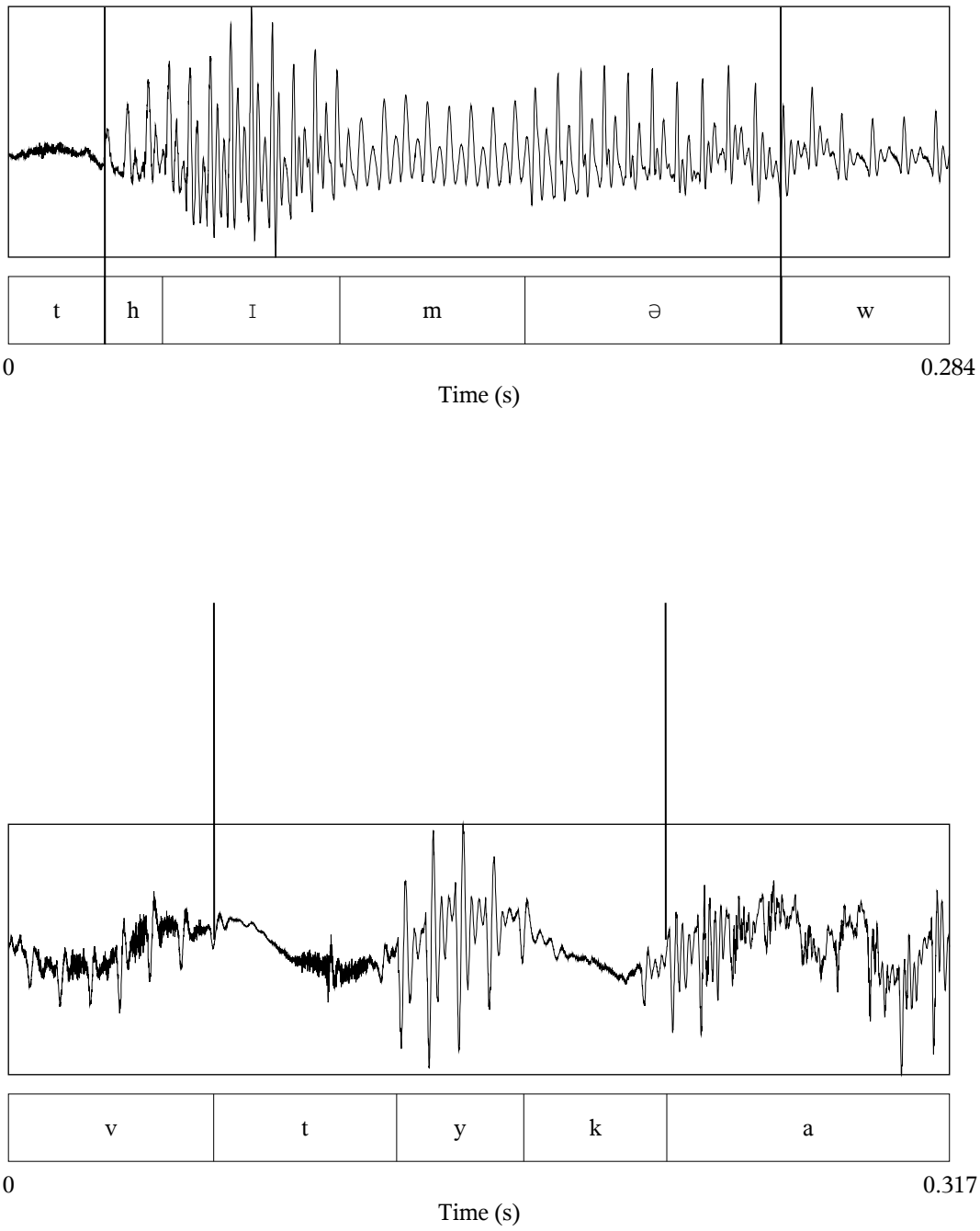

Figure 1: Examples of the segmentation for two stimuli in the experiment.

Since truncated speech often sounds very unnatural and may cause listeners to perceive an inserted labial or plosive consonant (e.g. Pols and Schouten, 1978), we added a 500 Hz square wave signal at the end of the preceding context. Square wave signals are not misperceived as

speech sounds (Warner, 1998), and are therefore suitable for the current purpose. The signal had a fixed duration (505 ms) and consisted of an onset with gradually increasing amplitude (5 ms), and a 500 ms part with a fixed amplitude of 52 dB. The overall intensity of each sound fragment (excluding the square wave) had an average of 70 dB.

4.3 Procedure

The basic procedure was identical to Experiment 1, the only differences being that the context was presented auditorily and that each speaker block was preceded by a brief familiarisation phase (mean: 19.8 seconds, range: 10.48 seconds to 36.99 seconds) that consisted of a short monologue by the speaker to introduce the participants to that speaker's (voice) characteristics, as in the control experiments. The participants listened to the speech via headphones. They heard each fragment twice (successively), because we wanted to establish whether listeners become more sensitive to subtle acoustic cues in the context when hearing the context a second time, given that the auditory materials contained high speech rates.

4.4 Results and discussion

The descriptive statistics for Experiment 2 show that participants, provided with four plausible options, selected the correct option in 39.47% of the cases. A regression model was fitted for response accuracy in the combined data set of Experiments 1 and 2. We included the fixed effect factors *Repetition*, *Position Correct*, and *Intelligibility in Isolation* (with the four quartiles as its levels, see Experiment 1), and, critically, whether the stimuli were presented orthographically or auditorily (henceforth *Presentation Mode*).

We found a main effect of *Presentation Mode* ($F(1,5995) = 10.45$, $p < 0.01$). Participants performed better if the preceding context was presented auditorily rather than orthographically (39.47% versus 33.27% correct). In addition to semantic/syntactic cues, listeners apparently used acoustic cues in the preceding context to predict the upcoming word, even though conversational speech is characterised by a high speech rate and contains reduced words, which may obscure

acoustic cues that have been proved to be useful in the comprehension of clear speech.

Possibly, participants relied on acoustic cues only because the semantic/syntactic cues in the context were insufficient. We investigated this hypothesis in Experiments 3 and 4. In these experiments, participants were presented with both the preceding and following context of the target words, again either orthographically or auditorily. We investigated whether the difference between the orthographic and auditory presentation mode established in Experiments 1 and 2 also holds when participants are provided with the preceding and following context.

5. Experiment 3

5.1 Participants

Twenty native speakers of Dutch from the pool of participants of the Max Planck Institute for Psycholinguistics were paid to take part in the experiment. None of the participants had taken part in any of the previous experiments, and none of them reported any hearing loss.

5.2 Materials

The context presented in Experiment 1 was extended to include words following the target word. Again, the target words themselves were not presented to the participants. In the presentation to the participants, the preceding and following context were separated by "_____", which indicated the position of the target word. The four options presented for each trial were identical to those in Experiment 1, except that the word that followed the target word was replaced by "_____", meaning that no word had been left out. See the example provided below.

Context:
Het geld is niet van jou en je staat _____ rood.
"The money is not yours and you are _____ in debt."

Options:

1. *altijd* "always" (correct, reduced word)
2. _____ (no words missing)
3. *bijvoorbeeld* "for instance"
4. *eigenlijk* "actually"

Since one of the four options was now replaced by "_____", which participants generally dispreferred, the percentages correct obtained in Experiments 3 and 4 cannot be well compared to those obtained in Experiments 1 and 2**,** but the experiments are comparable otherwise.

5.3 Procedure

The procedure was identical to that of Experiment 1, except that both the preceding and following context was presented on the screen, for ten seconds if it consisted of more than 25 words, otherwise for seven seconds.

5.4 Results and discussion

Participants selected the correct option in 37.27% of the trials. We analysed participants' selection of the correct answer versus the other options and included in the statistical model the fixed effects *Repetition*, *Position Correct*, and *Intelligibility in Isolation*. We did not find any significant effects.

If participants only used word frequency and N-gram frequency information to predict the missing target words, they would select the correct answer in maximally 33% of the trials, since only in these trials the correct answer had a higher frequency or N-gram frequency with the preceding and following word than the other options. Participants managed to perform significantly better than 33% correct (one-tailed t-test testing whether participants' performance was significantly higher than 33%: $t(19) = 2.82$, $p < 0.05$), which again suggests that participants

are sensitive to semantic/syntactic cues that are not captured by N-gram probabilities.

Since Experiment 2 showed that listeners can use acoustic cues in the preceding context to predict reduced tokens, they may also use acoustic cues if presented with both the preceding and following context. We conducted a fourth experiment, in which participants were presented auditorily with both the preceding and following contexts of the reduced tokens. If listeners also use acoustic cues if they are provided with the full context, they should perform better in this experiment than did the participants in Experiment 3.

6. Experiment 4

6.1 Participants

Twenty native speakers of Dutch from the same pool of participants as used in the previous experiments participated in the experiment for a small salary. None of these had taken part in these previous experiments, and none of them reported any hearing loss.

6.2 Materials

The auditory contexts presented in Experiment 2 were extended to include the words following the target words. Thus, participants were provided with the preceding context, the square wave, and then the following context. The target words themselves were not presented to the participants, although, as in Experiment 2, the context may contain (e.g. prosodic or spectral) traces of these target words. The four options were the same as those in Experiment 3 and were presented to the participants orthographically.

6.3 Procedure

The procedure was identical to that of Experiment 2.

6.4 Results and discussion

The descriptive statistics for Experiment 4 show that participants selected the correct answer in 48.03% of the trials. We fitted a regression model to compare the effects of the preceding and following auditory context to those of the preceding and following orthographic context (Experiment 3). We entered the predictors *Repetition*, *Position Correct*, *Intelligibility in Isolation*, and *Presentation Mode*. The results are provided in Table 2.

Predictor	F value	p value
Presentation Mode	16.58	< 0.01
Repetition	5.7	< 0.05
Position Correct	4.72	< 0.01
Presentation Mode * Repetition	5.39	< 0.05

Table 2: F values and significance values for the model comparing Experiments 3 and 4 (degrees of freedom: 5993).

The participants in the auditory context condition performed significantly better than those in the orthographic condition (48.03% versus 37.27% correct). Thus, even if provided with semantic/syntactic information in both the preceding and following context of reduced tokens, participants use acoustic cues in the context to improve their performance. Further, participants performed better after having heard the fragment a second time in the auditory presentation mode (50.73% versus 45.33% correct).

In conclusion, our results show that listeners use acoustic cues in the context if they are provided with the preceding semantic/syntactic context, but also if they are provided with the full semantic/syntactic context of reduced tokens. We now focus on the more specific research questions 1-3 formulated in the Introduction. Note that we investigated Question 4 already in the preceding sections.

7. Further analysis of combined results

Having established that listeners can use both semantic/syntactic and acoustic information in the context to predict reduced pronunciation variants, we now wish to determine their use of this information in more detail.

We first investigated whether longer contexts were more informative (and hence led to more correct responses) than shorter ones. The amount of context varied considerably between the speech fragments used in this study. The preceding context varied from 2 to 29 words, while the following context varied from 1 to 13 words. Longer contexts probably contain more acoustic and semantic/syntactic cues, and may hence increase the words' predictability.

With respect to just the acoustic cues, we investigated which properties may affect the likelihood that such cues inform listeners. Fast speech demands fast processing, and participants may therefore be less sensitive to acoustic cues in the context, the higher the speed rate (or, on the contrary more sensitive, since these are adverse listening conditions, as hypothesized by Hawkins and Smith, 2001). Further, transitional cues may be more informative and more salient in vowels than in consonants and therefore listeners may find it easier to predict upcoming words that are preceded by words ending in vowels.

With respect to the semantic/syntactic cues, we focused on the question whether N-gram probability effects were equally pervasive in the auditory (Experiments 2 and 4) and orthographic presentation modes (Experiments 1 and 3). Importantly, in the auditory experiments, the N-gram frequency information is in conflict with the acoustic cues, since we deliberately presented the participants with incorrect options that had higher N-gram frequencies with the words in the context than the correct answers in most trials. Participants presented with the auditory contexts may therefore have focused less on semantic/syntactic cues than the participants who did not have access to the conflicting acoustic information.

In order to address these questions, we fitted two regression models, the first one comparing Experiments 1 and 2 and the second comparing Experiments 3 and 4. With these models, we tested the roles of four types of variables.

First, we included variables concerning the length of the context. We tested for effects of the length of the preceding context (in both regression models) and the following context (only in the regression model for Experiments 3 and 4). Since the lengths of the preceding and following context did not show normal distributions, we also converted these numeric variables into two ordinal variables with four levels, representing their four quartiles (henceforth *Length of the Preceding Context* and *Length of the Following Context*). The quartile ranges for these two variables are presented in Table 3.

Level	Range in preceding context	Range in following context	Range in speech rate
Quartile 1	2-6 words	1-2 words	2.5 - 5.57
Quartile 2	7-9 words	3-5 words	5.57 - 6.54
Quartile 3	10-12 words	6-7 words	6.54 - 7.3
Quartile 4	13-29 words	8-13 words	7.3 – 9.3

Table 3: The quartile ranges for *Length of the Preceding Context*, *Length of the Following Context*, and *Speech Rate*.

Second, we included variables that may provide information about the likelihood that listeners can use acoustic cues given the characteristics of the speech signal. We tested for effects of *Speech Rate*, defined as the number of syllables of the phrase divided by the duration of the phrase (mean: 6.38 syllables per second). As *Speech Rate* was not distributed normally, we converted also this variable into an ordinal variable with four levels, representing the four quartiles. These quartile ranges are also presented in Table 3. We also incorporated as a predictor the type of the segment (i.e. consonant or vowel) immediately preceding the reduced word (henceforth *Preceding Sound*). Since both *Speech Rate* and *Preceding Sound* were only relevant

in the auditory modality, we expect that if these variables have an effect, these effects will interact with *Presentation Mode*.

Third, we included variables indicating the predictability of the correct answer based on the word's a-priori predictability (its unigram lexical frequency, range: 117.47-2752.85 per million) and its predictability given the two preceding words (word trigram frequency, range: 0-31.98 per million) or preceding and following word (henceforth *Surrounding Trigram Frequency*; only relevant for Experiments 3 and 4, range: 0-15.41 per million), relative to these same frequencies for the three other options. As the frequencies of the different options were not normally distributed, we could not calculate some relative continuous frequency measures and we therefore created three ordinal variables (*Word Frequency*, *Preceding Trigram Frequency*, and *Surrounding Trigram Frequency*) with three levels: Highest (i.e. the correct answer was the option with the highest frequency), intermediate, and lowest frequency (the correct answer had the lowest frequency). For the variables *Word Frequency* and *Preceding Trigram Frequency*, the intermediate level included the words with the second and third frequency rank. However, for the variable *Surrounding Trigram Frequency* we excluded the option that no word was missing, since for this option the *Surrounding Trigram Frequency* is the frequency of the preceding word with the two following words, while in some trials there was only a single word following the target. There were no strong correlations between these measures.

Finally, we also included as predictors the variables *Trial Number*, *Block Number*, and *Block Trial Number* (i.e. trial number within the given block), which can all capture effects of learning and/or fatigue. These predictors were included mostly in order to reduce the variance in the data.

To begin with, we fitted a regression model for Experiments 1 and 2. The final model is summarised in Table 4. We found a main effect of *Presentation Mode*: Participants gave more incorrect responses in the orthographic presentation mode (39.47% versus 33.27% correct). Further, we found a main effect of *Intelligibility in Isolation*: Participants made more errors for words that were difficult to recognise in isolation. This finding is unexpected given a listener driven account of speech reduction, which suggests that speakers reduce especially those words

that are highly predictable for the listener (Boersma, 1998). Further, we found an interaction between *Presentation Mode* and *Preceding Trigram Frequency*: In the orthographic presentation mode, participants had the tendency to choose an option with a relatively high trigram frequency (25.42% correct for target words with the lowest trigram frequency versus 35% and 33.64% for target words with the intermediate and highest trigram frequencies, respectively). The auditory experiment showed no frequency effects, which suggests that participants relied less on frequency information if they were also provided with acoustic cues in the context. We did not find any effects of *Preceding Sound*, *Length of the Preceding Context*, *Speech Rate*, *Repetition*, *Trial Number*, *Block Number*, or *Block Trial Number*.

Predictor	F value	p value
Presentation Mode	10.35	< 0.01
Intelligibility in Isolation	6.51	< 0.001
Position Correct	41.8	< 0.0001
Preceding Trigram Frequency	0.12	n.s.
Presentation Mode * Preceding Trigram Frequency	4.33	< 0.05

Table 4: F values and significance values for the model comparing Experiments 1 and 2 (degrees of freedom: 5988).

Subsequently, we fitted a regression model for Experiments 3 and 4, including the same predictors as for the correctness analysis of Experiments 1 and 2, in addition to *Surrounding Trigram Frequency*.

Predictor	F value	p value
Presentation Mode	16.41	< 0.001
Position Correct	4.66	< 0.01
Repetition	5.74	< 0.05
Surrounding Trigram Frequency	4.02	n.s.[6]
Presentation Mode * Repetition	5.32	< 0.05
Presentation Mode * Surrounding Trigram Frequency	3.66	< 0.05

Table 5: F values and significance values for the model comparing Experiments 3 and 4 (degrees of freedom: 5989).

The final model is summarised in Table 5. As also mentioned in the discussion of Experiment 4, participants gave more correct responses in the auditory than in the orthographic presentation

[6] This effect was only significant in the analysis of variance results; not in the summary of the model.

mode (48.03% versus 37.27% correct). Further, we found again an interaction of *Presentation Mode* with *Repetition*: In the auditory experiment, participants gave more correct responses after repetition (50.73% versus 45.33% correct). Apparently, listeners could make better use of the acoustic cues in the signal after repetition. More importantly for our research question, we found an interaction of *Presentation Mode* with *Surrounding Trigram Frequency*: In the orthographic presentation mode, participants gave more correct responses if the target word had the highest trigram frequency rather than the lowest trigram frequency with its preceding and following word (50.21% correct for targets with the highest trigram frequency versus 28.13% for targets with the lowest trigram frequency). No trigram frequency effects were present for the auditory presentation mode. This finding again suggests that participants relied less on these frequency cues if provided with acoustic cues in the context. Interestingly, neither presentation mode in Experiments 3 and 4 showed effects of *Preceding Trigram Frequency*, which suggests that participants shifted focus from the preceding to the preceding *and* following context. We did not find effects of *Preceding Sound*, *Length of the Preceding Context*, *Length of the Following Context*, *Speech Rate*, *Intelligibility in Isolation*, *Trial Number*, *Block Number*, or *Block Trial Number*.

In summary, with respect to research question 1, our data provide no evidence that participants are better in predicting reduced words if the context contains more words. Apparently, the words that we tested were equally predictable in their prosodic phrases, regardless of the lengths of these phrases. With respect to research question 2, our data provide no evidence that listeners are hindered by higher speech rate or benefit more from the transitional information in preceding vowels than in preceding consonants. This suggests that how good listeners are at interpreting acoustic cues in the context is largely independent of the precise phonetic characteristics of the prosodic phrases. Our data do provide more detailed information about how language users benefit from the semantic/syntactic context (research question 3). Whereas trigram frequency information is highly important in the absence of acoustic cues from the context, it plays only a marginal role if these acoustic cues are provided. Finally, with respect to research question 4, we found that the target word's reduction degree is correlated with its predictability if participants only have access to the preceding context. We will come back to this in the next section.

8. General discussion

Ernestus, Baayen, and Schreuder (2002) showed that contextual information is crucial for the understanding of reduced pronunciation variants. The present study investigated which contextual cues listeners can use to predict reduced word tokens or fixed expressions (henceforth *target words*, for the sake of convenience) in spontaneous speech. Participants were only presented with contextual information of the target words, and the target words themselves were always missing (the complete words in the orthographic transcriptions or those contiguous parts of the acoustic signals that contained cues that mostly pertained to the segments of the words). On the basis of the context, participants had to guess the missing target words, choosing from four semantically/syntactically plausible options (as judged by the authors) presented on the screen. We investigated the role of semantic/syntactic cues directly in the two orthographic experiments, in which the contexts of the reduced target words were presented in the form of orthographic transcriptions (Experiments 1 and 3).

First of all, we found that participants predicted the missing target words above chance in both the preceding and full context conditions (pure chance equalled 25%, since there were four options, and we obtained 33.27% and 37.27% correct for the two experiments). This finding is not self-evident, since we used low-predictability words, with low unigram and N-gram probabilities compared to the other three options presented on the screen. Our results therefore suggest that language models completely based on unigram or N-gram probabilities cannot explain language users' sensitivity to semantic/syntactic contextual information. Language users are sensitive to higher-level semantic/syntactic information as well.

In fact, we did not find any effects of word frequency on participants' response accuracy at all. This result is in line with previous findings showing that context reduces the effects of word frequency in visual word recognition (e.g. Becker, 1979; van Petten and Kutas, 1990; Rayner, Ashby, and Pollatsek, 2004). Van Petten and Kutas (1990) recorded ERPs while participants silently read semantically unrelated sentences. They found that low frequency words only

yielded larger event-related brain potentials than high frequency words if they appeared early in the sentences: The difference between high and low frequency words disappeared when sufficient context was available, and the words were predictable to some extent given the preceding words in the sentence.

In contrast, there was a reliable effect of trigram frequency. In the preceding context condition, participants were more likely to choose the correct answer if that word formed a relatively frequent word trigram with the two preceding words (i.e. if the word had the highest or an intermediate trigram frequency relative to the other three options). In the full context condition, participants were more likely to choose the correct option if the target formed the most frequent word trigram with the preceding and following word.

Our observation that, when provided with the full context, participants focused on the words' trigram frequency with the preceding *and* following words indicates that in addition to the preceding context, language users can use the following context to recognise words. This conclusion is supported by our finding that participants' accuracy is influenced by how intelligible the missing word is in isolation (established in a control experiment). Previous research suggests that speakers reduce especially those words that are highly predictable to the listener (e.g. Aylett and Turk, 2004). In contrast to this positive correlation, we found for the experiments with only the preceding context (Experiments 1 and 2, see Table 4) that participants made more errors for more reduced target words. This suggests that more reduced words were more difficult rather than easier to predict. This effect was absent when language users were also presented with the following context (Experiments 3), which shows that more reduced words were as predictable as less reduced words if also the following context was provided. Participants thus also extracted information from the following context in order to predict the missing words.

The effect of the following context on the predictability of reduced words in spontaneous speech is in line with previous research on the comprehension of laboratory speech (e.g. Warren and Warren, 1970; Warren and Sherman, 1974, Grosjean, 1985) and the comprehension of

spontaneous speech (e.g. Bard, Shillcock, and Altman, 1988). For example, in a study by Warren and Sherman (1974), listeners presented with sentences that contain deliberately mispronounced phonemes (e.g. *George waited for the deli[b]ery of his new color TV*) replaced by noise (leaving only misleading transitional cues) recover from the misleading acoustic cues based on the following context (i.e. listeners heard *deli[v]ery* instead of *deli[b]ery* in the example above). Our results indicate that the role of the following semantic/syntactic context generalises to the processing of reduced words, even if they have a low predictability (i.e. discourse markers/adverbs) and can be left out without significantly changing the sentences' meanings. As mentioned above, this result is not self-evident, as previous research showed that semantic priming from reduced words takes longer than from unreduced words (van de Ven, et al., 2011).

The literature contains roughly two accounts for how semantic/syntactic cues might facilitate word recognition. Van Petten and Kutas (1990) claim that a word with a high contextual predictability can be more easily integrated into the preceding context, which facilitates semantic processing. Alternatively, some researchers suggest that language users rely on contextual information to directly predict lexical items and narrow down their lexical search space (e.g. van Berkum, Brown, Zwitserlood, Kooijman, Hagoort, 2005). Both accounts can explain our data.

In the experiments in which the contexts were presented auditorily (Experiments 2 and 4), listeners could potentially use co-articulation cues because the four words they had to choose from differed in their word-initial sounds. Obviously, other acoustic cues (e.g. prosody) may have been useful as well. We found that participants were better at predicting the reduced words in the auditory experiments than in the orthographic experiments (43.75% versus 35.27% correct on average). Further, for listeners who heard the full context, these effects were larger when hearing a speech fragment a second time, which may be due to the large amount of information provided in this full context condition.

Participants thus use acoustic cues in the context to their advantage, which is no mean feat, since the provided contexts were extracted from spontaneous conversations and therefore had a high speech rate (mean: 6.38 syllables per second), included other reduced words, and showed high

variability (e.g. in speech rate, which varied between 2.5 to 9.3 syllables per second). Apparently, listeners can also use acoustic cues under these adverse listening conditions (as hypothesised by Hawkins and Smith, 2001). In fact, speech rate did not influence how well listeners predicted the missing words. We did not find effects of whether the sound preceding the missing word was a consonant or a vowel either. We did find effects of repetition: Participants performed better after hearing the auditory fragments a second time, whereas there was no improvement in the orthographic experiments. Further research is required to establish which acoustic cues in the context are particularly useful for the listener and, further, how exactly these cues facilitate the processing of spontaneous speech.

Several speech comprehension models assume a prelexical level of processing in which sounds are converted into abstract categories, such as phonemes (e.g. Shortlist B; Norris and McQueen, 2008). Shortlist B can account for listeners' sensitivity to acoustic cues in the context to the extent that these cues can be captured by the diphone database that is incorporated into the model. This database is based on careful speech and may therefore not contain acoustic transitional cues relevant for spontaneous speech. Further, previous research indicates that listeners are sensitive not only to acoustic cues in sounds directly preceding and following the target sound, but also to those that occur much earlier in the speech stream, for example acoustic traces of /r/ in syllables preceding /r/ (r-resonance, Kelly and Local, 1986), especially in spontaneous speech (e.g. Heinrich, Flory and Hawkins (2010)). Such acoustic cues cannot be captured by a diphone database but are likely to have influenced listeners in our experiments as well. These findings can more easily be explained by speech comprehension models that do not assume any pre-lexical abstraction, but that allow listeners to store all available acoustic cues in the input, for instance in the form of exemplars, to facilitate speech understanding (e.g. Polysp; Hawkins and Smith, 2001).

Importantly, our experiments also provide information on the relative contribution of semantic/syntactic and acoustic cues. These two types of cues can only be teased apart if there is a conflict between these two types of cues. Since the sentences were taken from natural speech, the acoustic properties of the contexts were always congruent with the target words. In most of

our stimuli (at least in 65% of the cases), these cues were in conflict with the immediately surrounding semantic/syntactic context, as the N-gram probabilities were low. We found that, when listeners are presented with acoustic contextual information that (frequently) conflicts with N-gram probability information, N-gram probabilities do not predict participants' choices at all. This suggests that if acoustic and semantic/syntactic cues in the context are in conflict, listeners consider acoustic cues more reliable than the N-gram probabilities of the words with their surrounding words. Further research has to show whether listeners also rely less on semantic/syntactic information if it is in general in agreement with the acoustic cues. If these results are replicated in further research, this interaction between the roles of semantic/syntactic and acoustic contextual information should be incorporated into current models of speech comprehension.

Finally, since listeners have great difficulty understanding highly reduced words in isolation, one may hypothesise that they use only context to understand such pronunciation variants. Thus, when English listeners hear *supposed to* pronounced as [səsə], they may deduce its meaning purely on the basis of the context. Our results are in contrast with this hypothesis. Participants performed above chance-level (more than 25% correct) in all four main experiments, but were unsuccessful at predicting the reduced words across the board (less than 50% correct). Nevertheless, one of our control experiments showed that listeners can well recognise the target words when they hear the preceding and following auditory context together with these words (more than 90% correct). We therefore conclude that listeners need not only the context but also the reduced word itself to comprehend this word, which underlines the significance of the (albeit reduced) acoustic properties of highly reduced pronunciation variants.

In conclusion, the present study investigated the contribution of the various types of contextual information available in spontaneous speech. Whereas most studies focus on the role of context in the comprehension of content words in simple sentences in laboratory speech, the present study investigated to what extent listeners can use context to process low predictability reduced words (e.g. discourse markers and adverbs) in natural spontaneous speech. Our results show that listeners use both the preceding and following context to process such words, and that they are

sensitive to semantic/syntactic as well as acoustic cues in spontaneous speech contexts, but that they favour acoustic cues in the case of a conflict.

References

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

Bard, E.G., Shillcock, R.C., & Altmann, G.T.M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*, 44, 395–408.

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance* 5, 252–259.

Bell, A., Brenier, J., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60, 92-111.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., and Gildea, D. (1999). Forms of English function words – effects of disfluencies, turn position, age and sex, and predictability. *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco, California, 395–398.

Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam, Amsterdam, The Netherlands.

Bybee, J. L. (2001). *Phonology and language use*. Cambridge University Press, Cambridge.

Davis, M. H., Marslen-Wilson, W. D., and Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition.

Journal of Experimental Psychology: Human Perception and Performance, 28, 218-244.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch: A Corpus-Based Study of the Phonology-Phonetics Interface*. Utrecht: LOT.

Ernestus, M. and Baayen, R. H. (2011). Corpora and exemplars in phonology. In J. A. Goldsmith, J. Riggle, and A. C. Yu (Eds.), The handbook of phonological theory (2nd ed.) (pp. 374-400). Oxford: Wiley-Blackwell.

Ernestus, M., Baayen, R. H., and Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language* 81, 162–173.

Ernestus, M. and Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics* 39, 253-260.

Goldsmith, J. (1976). An overview of autosegmental phonology. *Linguistic Analysis*, 2, 23-68.

Gow Jr., D. (2002). Does English Coronal Place Assimilation Create Lexical Ambiguity? *Journal of Experimental Psychology: Human Perception and Performance* 28, 163-179.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications, *Perception and Psychophysics* 38(4), 299–310.

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

Hawkins, S. and Smith, R. (2001). Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics – Rivista di Linguistica* 13, 99-188.

Hawkins, S. and Warren, P. (1994). Phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics* 22, 493–511.

Heinrich, A., Flory, Y., and Hawkins, S. 2010. Influence of English r-resonances on

intelligibility of speech in noise for native English and German listeners. *Speech Communication* 52, 1038-1055.

Hooper, J. B. (1976). An introduction to natural generative phonology. Academic Press, INC, New York.

Howes, D.H. (1954). On the interpretation of word frequency as a variable affecting speech of recognition. *Journal of Experimental Psychology*, 48, 106-112.

Howes, D.H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, 29, 296-305.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59 (4), 434–446.

Janse, E. and Ernestus, M. (2011). The roles of bottom-up and top-down information in the recognition of reduced speech: evidence from listeners with normal and impaired hearing. *Journal of Phonetics* 39, 330-343.

Johnson, K. (2004). Massive reduction in conversational American English. In: Spontaneous speech: data and analysis. *Proceedings of the 1st session of the 10th international symposium*. The National International Institute for Japanese Language, Tokyo, Japan, 29–54.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., and Raymond, W. (1998). Reduction of English function words in switchboard. *Proceedings of ICSLP-98*, 3111–3114.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In: Bybee J., Hopper, P. (Eds.), *Frequency and the emergence of linguistic structure*. John Benjamins, Amsterdam, 229–254.

Kelly, J. and Local, J.K. (1986). Long-domain resonance patterns in English. In: *Proceedings of IEE International Conference on Speech Input/Output: Techniques and Applications*, London, pp. 304-309.

Marslen-Wilson, W. and Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review* 101, 653–675.

Martin, J. and Bunnell, T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 473-488.

McDonald, S. and Shillcock, R. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science* 14(6), 648–652.

Mitterer, H. and Ernestus, M. (2006). Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch. *Journal of Phonetics* 34, 73-103.

Morton, J. and Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior* 15, 43–51.

Newbigging, P.L. (1961). The perceptual reintegration of frequent and infrequent words. Canadian Journal of Psychology, 15, 123-132.

Nooteboom, S. G. (1972). *Production and perception of vowel duration: A study of the durational properties of vowels in Dutch*. PhD thesis, University of Utrecht, Utrecht, The Netherlands.

Nooteboom, S. G. and Doodeman, G. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America* 67, 276–287.

Norris, D. and McQueen, J.M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357-395.

Oostdijk, N. (2002). The Design of the Spoken Dutch Corpus. In P. Peters, P. Collins and A. Smith (eds.): New Frontiers of Corpus Research. 105-112. Amsterdam: Rodopi.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech* (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State
University (Distributor).

Pluymaekers, M., Ernestus, M., and Baayen, R. (2005). Lexical frequency and acoustic reduction in spoken Dutch, *Journal of the Acoustical Society of America* 118(4), 2561-2569.

Pols, L. C. W. and Schouten, M. E. H. (1978). Identification of deleted consonants, *Journal of the Acoustical Society of America* 64, 1333-1337.

Raymond, B., Makashay, M., Dautricourt, R., Johnson, K., and Pitt, M. (2001). *An introduction to the Buckeye speech corpus.* Poster presented at the 142nd meeting of the Acoustical Society of America, Ft. Lauderdale, FL (Dec).

Rayner, K., Ashby, J. and Pollatsek, A. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the e-z reader model.
Journal of Experimental Psychology: Human Perception and Performance
30, 720–732.

Salverda, A., Dahan, D., and McQueen, J. (2003). The role of prosodic boundaries
in the resolution of lexical embedding in speech comprehension. *Cognition*
90, 51–89.

Savin, H.B. (1963). Word-frequency effect and errors in the perception of speech. Journal of the

Acoustical Society of America, 35, 200-206.

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). E-prime users guide.

Soloman, R.L. and Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. Journal of Experimental Psychology, 43, 195-201.

van Berkum J. J. A., Brown C. M., Zwitserlood P., Kooijman V., and Hagoort P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, 443-467.

van de Ven, M.A.M., Tucker, B.V., and Ernestus, M. (2011). Semantic context effects in the comprehension of reduced pronunciation variants. *Memory & Cognition* 39, 1301-1316.

van den Brink, D., Brown, C., and Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 and N400 effects. *Journal of Cognitive Neuroscience* 13, 967-985.

van den Brink D., Brown C., and Hagoort, P. (2006). The cascaded nature of lexical selection and integration in auditory sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32, 364-372.

van Petten, C. and Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition* 18, 380–393.

Warner, N. (1998). *The Role of Dynamic Cues in Speech Perception, Spoken Word Recognition, and Phonological Universals*. PhD thesis, University of California, Berkeley.

Warner, N., Fountain, A., and Tucker, B.V. (2009). Cues to perception of reduced flaps. *Journal of the Acoustical Society of America* 125, 3317-3327.

Warren, R. and Sherman, G. (1974). Phonemic restorations based on subsequent context, *Perception and Psychophysics* 16, 150–156.

Warren, R. & Warren R. (1970) Auditory illusions and confusions. *Scientific American*, 223 (6), 30-36.

West, P. 1999. Perception of distributed coarticulatory properties of English /l/ and /r/. *Journal of Phonetics* Vol. 27(4), 405-426.

West, P. 2000. Long-distance coarticulatory effects of British English liquids: an EMA, EPG and Acoustic study. *Oxford University Working Papers in Linguistics Philology and Phonetics.* 5, 73-86.

Wiggers, P. (2008). *Modelling context in automatic speech recognition*. PhD thesis, Delft University of Technology, Delft.

Zipf, G. K. (1935). The Psycho-Biology of Language. Houghton Mifflin, Boston.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition* 32, 25–64.