GLOSSA psycholinguistics

Cohen, C. (2024). Predicting this rock: Listeners use redundant phonetic information in online morphosyntactic processing. *Glossa Psycholinguistics*, 3(1): 24, pp. 1–52. DOI: https://doi.org/10.5070/G60111493

eScholarship University of California

Predicting this rock: Listeners use redundant phonetic information in online morphosyntactic processing

Clara Cohen, University of Glasgow, UK, clara.cohen@glasgow.ac.uk

Pronunciation variation is systematic, and provides listeners with cues to what the speaker is about to say. Shortened stems, for example, can indicate an upcoming suffix, while lengthened ones can indicate a word boundary follows. Previous work has shown that listeners draw on these cues to distinguish polysyllabic words, like *rocket*, from monosyllabic words, like *rock*. This strategy is useful in morphological processing, as additional morphological structure often adds additional syllables. The current study asks (i) whether listeners use these cues to distinguish words that differ only in morphological structure with no change in syllable count (e.g., *rock*/*rocks*); and (ii) how surrounding morphosyntactic context affects listeners' ability to use these cues. *Ideal observer* models predict that listeners should be attentive to phonetic detail in all contexts regardless of how much new information it offers, while the *strategic listener* account allows listeners to dynamically adjust their attentiveness to phonetic detail based on its information value in context. In a visual-world eye-tracking study, English-speaking listeners were presented with utterances containing target nouns whose stem durations were manipulated to provide cues to the presence or absence of (a) a plural suffix (*rock* vs. *rocks*) or (b) a second, non-morphological syllable (*rock* vs. *rocket*). These words were embedded in two contexts: (i) preceded by agreeing determiners, which rendered stem duration cues redundant for predicting the presence or absence of a suffix (*this rock*/*these rocks*), and (ii) preceded by non-agreeing determiners (*the rock(s)*), where stem duration cues carried more information. The results are consistent with ideal observer models: listeners are highly attentive to all acoustic detail, and especially so when it is predictable (and hence redundant), as long as they have the cognitive resources to handle it.

1. Introduction

Speech is full of temporary ambiguities that can pose problems for real-time incremental perception. These ambiguities are ubiquitous at the level of the individual word, where multiple lexical items might start with the same sequence of syllables (e.g., *beetle* and *beaker*; Allopenna et al., 1998), or one word might sit inside another, as *rock* sits inside *rocket* and *pan* sits inside *pansy*. Listeners are highly attentive to subphonemic phonetic cues that distinguish these temporarily ambiguous sequences from each other when they represent distinct lexical items (Blazej & Cohen-Goldberg, 2015; Davis et al., 2002; Salverda et al., 2003; Shatzman & McQueen, 2006; R. Smith & Rathcke, 2017). The evidence is more scanty when it comes to distinguishing inflectional forms from each other, and the research that has explored it has tended to focus on words in isolation (Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005). Are listeners equally attentive to these same cues when distinguishing between inflectionally related words in sentence contexts?

On the one hand, why would they not be? If listeners are capable of using informative acoustic-phonetic cues to aid online speech perception, it seems foolish to disregard them. Yet, on the other hand, if those acoustic-phonetic cues are redundant in the broader morphosyntactic context, then it might be equally foolish to waste working memory to attend to them.

The goal of this work is to explore how far listeners go in attending to subphonemic cues to word structure. Are they as attentive to the fine phonetic detail that distinguishes inflectionally related forms as they are to the detail that distinguishes temporarily ambiguous lexical items? And if so, then does the attention to detail vary depending on its informativeness or redundancy in the surrounding context?

1.1 Subphonemic durational cues in production

Let us consider first the nature of subphonemic cues to word structure that are available to listeners. One of the most commonly studied set of features relates to duration. As a general rule, it is well known that adding more phonological material to a word or syllable tends to shorten the absolute duration of the individual components. This general phenomenon can be understood as breaking down into two patterns: polysyllabic shortening and segmental compression.

Under *polysyllabic shortening*, the identical set of segments tends to be produced with shorter duration when more syllables are added to a word. For example, sequences like [kæp] and [hæm] are shorter when they form the first syllable of a word like *captain* or *hamster* than when they form a single-syllable word like *cap* or *ham* (Davis et al., 2002; Salverda et al., 2003).[1]

[1] This pattern is not universal: In languages which display it, such as English and German, it favors accented words (Siddins et al., 2013; White & Turk, 2010), and in languages that use duration as a phonological feature, such as Finnish, it is absent entirely (Suomi, 2007).

When words acquire additional morphological structure, they often increase their syllable count, and so become subject to polysyllabic shortening. Thus, the stem *speed* is longest when it is a free-standing word, shorter when it is the first syllable followed by a suffix, such as *speedy*, and shorter still when further suffixes add a third syllable, as in *speediness* or *speedily* (Lehiste, 1972; but see also Blazej & Cohen-Goldberg, 2015; Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005).

Yet adding morphological structure does not always increase syllable count. Singular and plural nouns in English, which are related by morphological suffixation, typically suffixed with the single-segment suffix *-s*, escape the conditions for polysyllabic shortening, because the plural *-s* suffix does not always add additional syllables to the stem. Singular *rock* and plural *rocks* in English are both one syllable. Nevertheless, even one-syllable words contain durational cues to the presence of an upcoming suffix, in a pattern known as *segmental compression*.

Segmental compression in English has been reported for decades. Klatt (1975), for example, observed that consonants in syllable onsets were shorter when they were part of a consonant cluster than when they were a simple onset. Katz (2012) observed that vowels are shorter when they occur in syllables with a coda than syllables with no coda, and that vowels in syllables with simple codas shorten further when the coda becomes complex, although only for certain types of consonants. Munhall et al. (1992) also observed significant shortening of vowels when syllable codas became more complex – including cases in which syllable codas were made more complex by the addition of a final *-s*. And most recently, Cohen and Carlson (2024) observed in a corpus study that both noun stems and especially verb stems shorten when a suffixal *-s* is added. These findings lead us to expect that adding a plural suffix to a noun stem should shorten the stem, providing listeners with a durational cue signaling the upcoming suffix.

1.2 Subphonemic durational cues in perception

Are these durational cues actually used to guide perception? For polysyllabic shortening, a large body of research seems to indicate that listeners are robustly sensitive to them. In a series of gating and priming experiments, Davis et al. (2002) found that when English-speaking listeners were presented with shortened initial syllables (e.g., *cap*) they expected that the rest of the word would form a polysyllabic continuation (e.g., *captain*). Salverda et al. (2003) replicated these findings for Dutch listeners using visual-world eye-tracking. Listeners were presented with a screen containing an image of a single-syllable word (*ham*) and its corresponding polysyllabic counterpart (*hamster*). Listeners then heard sentences in which the polysyllabic target (*hamster*) contained a manipulated initial syllable. This syllable had been spliced in from other recorded tokens of either *ham* or *hamster*, carefully selected so that its duration was either long or short. Regardless of whether the spliced syllable came from *ham* or *hamster*, listeners looked more towards the target, *hamster*, when the initial syllable had shorter duration, signaling a longer

polysyllabic word, than when it had longer duration, signaling a shorter free-standing word. In other words, listeners showed that they rely on the duration of the initial syllable to distinguish single-syllable nouns from polysyllables.

Shatzman & McQueen (2006) extended Salverda et al.'s (2003) study to show that Dutch listeners' ability to use this polysyllabic shortening pattern in perception reflects a highly generalized knowledge of the phenomenon. In their study, all target words were novel words (e.g., *bap* and *baptoe*), which listeners had been trained to associate with novel shapes before the eye-tracking task. Crucially, the recordings used to teach listeners the word-shape associations during the training task had been manipulated to neutralize all polysyllabic shortening patterns. The duration of [bæp] was identical in the training recordings of both *bap* and *baptoe*. During the test phase of the experiment, the durations of the initial syllables were then lengthened or shortened, producing durational distinctions that the listeners had not encountered when learning the words. Yet despite their lack of experience with polysyllabic shortening in these particular words, listeners nevertheless looked more towards the *bap* image if the duration of the stem was long, and more towards the *baptoe* image if the duration was shortened. The results suggest that the polysyllabic shortening patterns that listeners have learned throughout a lifetime of exposure generalize strongly enough to overrule the durational neutrality of the novel training words.

This ability to generalize extends to morphologically related words. Kemps and colleagues (Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005) used number decision and lexical decision tasks to study listeners' perception of English and Dutch words with inflectional or derivational suffixes, all of which added an additional syllable to a word stem. They then manipulated these words, so that originally unsuffixed words were presented either unsuffixed, or with a suffix spliced on, while originally suffixed words were presented with the suffix removed, or with a suffix present. They observed that reaction times were slowed in both English and Dutch listeners when the duration of a stem did not match the constructed word's structure. Finally, Blazej & Cohen-Goldberg (2015) directly manipulated the duration of the initial syllables of single-syllable English stems (e.g., *clue*) and morphologically complex derivations (e.g., *clueless*). Focusing on the single-syllable stems as their critical targets, they used acoustic analysis software to explicitly set the duration of single-syllable targets (*clue*), so that the duration ratio of shortened targets to lengthened targets mirrored the ratio of initial syllables in suffixed words to free-standing words in their naturally recorded stimuli. They observed that listeners' mouse tracks veered more towards the suffixed competitor (*clueless*) when the free-standing target (*clue*) had a shortened syllable than when it had a lengthened syllable.

The findings summarized above all present a picture showing that, at least as far as polysyllabic shortening is concerned, listeners are certainly capable of perceiving and using durational cues to word structure to guide their speech perception. After long initial syllables, listeners expect word boundaries; and after shortened initial syllables, they expect more of the word to follow, whether that additional word structure is a suffix or a continuation of the same root.

What happens, then, when the addition of a suffix does *not* add an additional syllable? Can listeners draw on segmental compression patterns to more quickly distinguish a single-syllable, unsuffixed noun, like *rock*, from a plural, like *rocks*? If so, then listeners should perform better when distinguishing singular nouns from plural competitors when the singular nouns have lengthened duration, compared to when those same singular nouns have been shortened.

The first goal of this study, therefore, is to confirm that listeners can indeed draw on segmental compression patterns in the same way they draw on polysyllabic shortening. This question is particularly interesting because inflectionally related words, like singulars and plurals, differ from pairs like *rock/rocket* and *clue/clueless* in two key ways. First, they belong to the same lexeme, and therefore share highly similar sets of lexico-semantic features. The consequence of misinterpreting a singular for a plural would lead to less communicative chaos than misinterpreting a singular for an unrelated two-syllable word. Looking for a *rock* instead of some *rocks* would still lead eventually to an identification of a mostly-correct target. Looking for a *rock* instead of a *rocket* will not. Listeners may therefore have learned that attending to detail that distinguishes different lexemes (*rock/rocket*) avoids more miscommunication than attending to detail distinguishing different word forms of the same lexeme (*rock/rocks*). If that is the case, then segmental compression effects on perception may be weaker or entirely absent, compared to polysyllabic shortening effects.

The second goal of this study capitalizes on another key feature of inflectionally related pairs: their morphosyntactic connections to the rest of the sentence context. The number of a noun can be highly predictable in some contexts – e.g., following agreeing determiners – and entirely unpredictable in others. Therefore, if listeners do draw on segmental compression patterns to aid perception, the extent to which they use them may vary as a function of the predictability of a given noun's number in its morphosyntactic context. There are three possible accounts for how listeners might integrate morphosyntax with subphonemic phonetic detail. I shall call these accounts the *baseline ideal observer* account; the *phonetic predictor* account; and the *strategic listener* account. All three accounts allow for listeners to draw on durational detail to inform their differentiation of singular, plural, and two-syllable words. They differ, however, in their predictions of how morphosyntactic context affects listeners' use of those cues.

1.3 Baseline ideal observers

Ideal observer models propose that listeners incorporate all possible sources of information to estimate the most likely interpretation of a target utterance (Kleinschmidt & Jaeger, 2015; Norris & McQueen, 2008). Such models are probabilistic: Listeners do not fully settle on one predicted interpretation as the utterance unfolds (e.g., "singular noun"), but rather derive probability distributions over possible interpretations (e.g., "singular = 73%, plural = 27%"), updating them as new information becomes available in the incoming input. Whether this new input takes the form of long-term repeated, accumulated exposure or the short-term immediate

speech context, the mechanism is the same: When the input is highly consistent with a particular interpretation (e.g., an unsuffixed noun stem), the probability of that interpretation is increased, while the probability of competing interpretations is down-weighted.

This view of incremental, probabilistic speech processing has much in common with earlier perspectives, such as the polysystemic speech perception (Polysp) view introduced by Hawkins (2003). It differs, however, by focusing on listeners' abilities to identify linguistic structures, rather than aiming purely at extracting meaning.[2] It also employs an explicitly Bayesian framing of its computational mechanics. According to this Bayesian architecture, listeners' expectations form prior distributions, which, upon the arrival of new information, are updated to form posterior distributions over likely utterances. Regardless of the shape of the prior distribution, new information will always combine with it in the same way. Thus, even if previous portions of the utterance have provided sufficient information to effectively disambiguate an acoustic signal, a listener will continue to integrate new input to update the posterior distribution. In cases where the previous input identifies the correct interpretation to a high degree of certainty, the new input is informationally redundant. "Redundant" here does not mean irrelevant or useless, but rather refers to the amount of information that a given input can carry. In contexts where previous input can be used near-categorically to identify the intended interpretation, then further input pointing in the same direction carries no new information. Equivalently, from a Bayesian perspective we can say that this new input can do little to affect the shape of the listener's posterior distribution. The distribution is already so heavily skewed towards the correct interpretation that skewing it further has very little practical effect on the listener's decisions. Nevertheless, the listener does not disregard this input. Regardless of its informational redundancy, it is allowed to influence the shape of the posterior, and in the event that the speaker misspoke, or the listener misheard, this strategy allows the listener to recover as the conflicting new information alerts them to the error.

This view of speech perception is "greedy": Listeners attend to *all* possible sources of information to ensure that their predicted probability distributions over upcoming material are as detailed as possible. Such use of cues is consistent with research showing that listeners form and store highly detailed representations of previously encountered linguistic structures (Goldinger, 1998; Hawkins, 2003; Johnson, 1997, 2007; Pierrehumbert, 2002), down to the individual patterns used by different speakers (Nygaard & Pisoni, 1998; Nygaard et al., 1994; L. B. Smith & Colunga, 2012; R. Smith, 2015).

To illustrate how this account applies to the test case here, of recognizing singular or plural nouns in different contexts, consider (1)–(2) below, with particular attention to the underlined noun phrase.

---

[2] The assumption that listeners focus primarily on identifying structures rather than meaning is challenged by recent work, which has found that pronunciation patterns may well be driven more by semantic associations than by features traditionally associated with discrete structures, such as word frequency or morphological composition (Gahl & Baayen, in press; Saito et al., 2024).

(1) a. It seems that the rock attracts migrating songbirds.
b. It seems that the rocks attract migrating songbirds.

(2) a. It seems that this rock attracts migrating songbirds.
b. It seems that these rocks attract migrating songbirds

Both (1a) and (2a) feature a singular noun, *rock*, while (1b) and (2b) feature a plural noun, *rocks*. But the two pairs differ in the quantity and nature of cues pointing to the noun's number up to the moment of the suffix onset. In (1a–1b), the cues are minimal and found in the domain of whichever coarticulatory processes go along with the the presence or absence of a final plural suffix. Such cues for (1b), for example, might comprise an adjustment to the place of articulation in the /k/ to accommodate the upcoming alveolar /s/, along with a slight shortening of the noun stem duration due to segmental compression (Katz, 2012; Klatt, 1975; Munhall et al., 1992). For (1a), a cue might be the corresponding lengthening of the segments that would result from the lack of any additional coda consonant. By contrast, (2a–2b) not only contain the same acoustic information in the noun stem as (1a–1b), but also add high-level morphosyntactic cues to upcoming plurality in the preceding demonstrative determiners *this* and *these*. Under an ideal observer model, the listener will therefore have a weaker probabilistic expectation about the number of the noun in (1a–1b), formed solely on the basis of the acoustic information in the noun stem, than in (2a–2b), where the prediction is supported by the combination of acoustic and morphosyntactic cues in the preceding words.

Crucially, although the ideal observer model might predict stronger overall expectations in the presence of multiple cues, as in (2a–2b), it does not predict different *use* of those cues. A listener who already has strong probabilistic expectations of a singular noun from the determiner in (2a) will still be able to strengthen those expectations further when they hear the patterns of lengthened stem duration in the noun stem. Likewise, a listener who has weaker expectations, due to the non-agreeing determiner *the* in (1a), will also be able to adjust those expectations when they encounter the same lengthening in the noun stem. Thus, ideal observer models predict an *additive*, rather than interactive, combination of cues: Regardless of the presence of morphosyntactic information, subphonemic detail will still be incorporated in the same way to the listener's perceptual processing. Listeners may be slower to process the final noun in sentences like (1) than sentences like (2), because of a more diffuse probability distribution, but they will derive equivalent benefit from acoustic information that matches the intended interpretation. This idea – of additive rather than interactive benefits of additional cues to noun number – is the *baseline ideal observer* account.

1.4 Phonetic predictors

A variant account of ideal listener models, which I shall call the *phonetic predictor* account, argues that listeners not only use these durational cues even when they are redundant in context, but

actually predict them from that same context. This view is built upon two sets of evidence. First, listeners are adaptable in their use of phonetic information not just in response to accumulated exposure (Mitterer & Reinisch, 2017; Norris et al., 2003; Nygaard & Pisoni, 1998; L. B. Smith & Colunga, 2012), but also in response to the immediate speech context. Second, listeners can use perceptual processing *predictively*, such that their ability to process incoming speech is assisted when the phonetic realization of that speech matches predictions formed from the preceding context.

Evidence regarding adaptability in context is plentiful. Dilley and Pitt (2010), for example, observed that listeners use speech rate in deciding whether highly reducible function words, such as *or* or *are*, are present in the speech stream. The identical acoustic stimulus is heard as, e.g., *leisure or time* when the preceding speech rate is fast, but as *leisure time* when the preceding speech is slowed down. Phoneme category boundaries are even more changeable, shifting as a function of the lexical context of the embedding word (Ganong, 1980); the semantic context of the embedding sentence (Borsky et al., 1998); the pragmatic context of the discourse actors (Rohde & Ettlinger, 2012); and, crucially, the syntactic context of the target phoneme. van Alphen and McQueen (2001), for example, reported that Dutch listeners shift their interpretation of an ambiguous plosive on a /t-d/ continuum as a function of the preceding syntactic context. When the context leads participants to expect a noun phrase, they are more likely to report the plosive to be /d/, resulting in the determiner *de*, which can appropriately precede nouns. When context leads participants to expect a verb phrase, they shift the category boundary and report more /t/s, resulting in the infinitival marker *te*. Fox & Blumstein (2016) found a similar pattern in English one word further downstream. They asked listeners to distinguish between an ambiguous bilabial plosive that could form a noun in one case and a verb in the other (e.g., [?aɪ] allows the noun *pie* if the plosive is interpreted as voiceless, and the verb *buy* if the plosive is interpreted as voiced). In contexts requiring a noun, participants shifted their boundary to produce more noun judgments (e.g., reporting more [p], to produce *pie*). In contexts requiring a verb, participants shifted their boundary in the other direction, (e.g., reporting more [b], resulting in *buy*).

This type of adaptation applies to morphological structures as well as phoneme category boundaries. Barden & Hawkins (2014) exposed English-speaking participants to an unfamiliar accent that produced the prefix *re-*, typically pronounced [ɹiː], instead as [ɹɪ]. After exposure, listeners completed a speech-in-noise task. Those who had been exposed to the novel accent more accurately identified new words containing non-standard [ɹɪ] than members of the control group, who had not previously heard the accent. Importantly, this perceptual learning was stronger when the non-standard [ɹɪ] formed a prefix, such as *re-supply*, than when it formed a non-morphological initial syllable, such as *recent*. Thus, the adaptation cannot be interpreted solely as a category boundary shift at the phoneme or syllable level, but also reflects listeners' recognition of the prefixal status of the *re-* syllable used during training.

In sum, it is undeniable that listeners can flexibly use both syntactic and morphological context to inform their processing of phonetic detail. Yet the studies described above relied on tasks such as transcription and categorization of stimuli. They capture listeners' offline judgments about the nature of the stimuli, rather than their real-time processing of what they hear. The second foundation of the phonetic predictor account is that listeners use perceptual processing *predictively*, such that they draw on their knowledge of fine phonetic detail to guide their real-time processing of incoming input in the speech stream.

For evidence of this ability, we must look to work on accent adaptation. Trude & Brown-Schmidt (2012), for example, used visual-world eye-tracking to examine listeners' real-time use of talker identity to guide perception of key words, both in isolation (Experiments 1 and 2a) and in a sentence context (Experiment 2b). Listeners were exposed to one of two talkers, one male and one female, during a training phase. The male talker used a regional variety of American English that raises /ae/ to [eɪ] before /g/, while the female talker used a General American accent. Listeners were asked to distinguish target words like *tack* from competitors like *tag*. They were faster to look to the target when listening to the regional-accented male talker than the standard-accented female talker, and especially so when they had information about the talker's identity – either through a visual cue of the talker's face when the target was an isolated word (Experiment 2a) or through a preceding sentence preamble that made the talker's identity clear (Experiment 2b). In other words, listeners learned that the regional-accented talker would have raised the /æ/ vowel to [eɪ] if the target ended in /g/. With the standard-accented talker, listeners knew not to expect such raising, and so had to wait longer for the disambiguating final consonant to identify the target. Thus, this study shows that listeners can form highly specific predictions about the phonetic form that upcoming words might take, and use those predictions in real time to guide perception.

Yet there are limits on this type of phonetically detailed prediction. Trude et al. (2013), for example, failed to replicate this pattern when listeners were exposed to a French-accented speaker who shifted /i/ toward [ɪ] before voiceless plosives, while maintaining the [i] vowel quality before fricatives. When asked to distinguish a target word like *wheat* from a competitor like *wheeze*, listeners had more difficulty with the French-accented talker – even though that talker distinguished between the vowel qualities in these two words – than with the English-accented talker, who used the same [iː] in both target and competitor. In this case, listeners could not use this information predictively.

These studies suggest that listeners are capable of adapting their processing of phonetic detail across a wide variety of contexts, and that they can develop quite detailed and specific phonetic predictions in at least some situations. Nevertheless, such findings have only been demonstrated in relation to the phonetic realization of a specific (morpho)-phonological contrast, such as the difference between two realizations of a prefix (Barden & Hawkins, 2014) or a vowel (Trude &

Brown-Schmidt, 2012; Trude et al., 2013); a place of articulation contrast in fricatives (Rohde & Ettlinger, 2012); or a voicing contrast in plosives (Borsky et al., 1998; Fox & Blumstein, 2016; Ganong, 1980; van Alphen & McQueen, 2001). Lengthening or shortening a stem in the absence or presence of a suffix is not a phonological process of the type these perceptual learning studies have examined. There is no categorical contrast to be realized. It is a much more gradient phonetic process, and so it remains an open question whether listeners include that degree of detail in their predictions.

If they do, then we can expect a different set of results from the baseline ideal observer account summarized in 1.3. Consider once again the contrast between (2a) and (1a). In the case of (2a), the presence of an agreeing determiner signals that the upcoming noun is singular, and hence likely to be unsuffixed. Unsuffixed nouns tend to have longer stem durations. Thus, the determiner could allow the listener to adjust their probabilistic distribution over likely stem realizations to favor longer duration. When the listener encounters a noun stem that is, in fact, lengthened, it will land in a high-density region of this probability distribution, allowing the stem to be more quickly processed. If the noun stem is shortened, it will land in the lower-density tail of that distribution, and so take longer to be processed. As a result, after a singular determiner, a lengthened singular stem should be processed more easily than a shortened singular stem.

By contrast, after non-agreeing determiners, as in (1a), no predictions about noun number, suffix status, and hence stem duration, can be formed. The probabilistic distribution over likely stem durations will need to remain flatter, to encompass a wider range of durations. The probability density of the distribution where a lengthened stem lands will be more similar to the density in the region where the shortened stem lands. The listener will therefore process the two stem durations more similarly after non-agreeing determiners than after agreeing determiners. They should also take slightly longer to make a decision, because they will need to wait until they reach the end of the word to incorporate morphosyntactic information into their distribution of likely interpretations.

In sum, ideal observer models predict two possible sets of outcomes. The phonetic predictor account says that listeners make phonetically detailed predictions about stem durations, and so the benefits of listening to nouns whose stem durations match their number should be stronger after agreeing determiners, which allow those durations to be predicted, than after non-agreeing determiners. In other words, it predicts an *interactive* use of cues. This contrasts, recall, with the baseline ideal observer account, which predicts an *additive* use of cues. Since listeners' predictions under the baseline ideal observer account do not include as high a level of phonetic detail, the benefits of listening to nouns with expected stem durations will not differ across determiner types.

Note that both of these accounts agree in granting listeners equal ability to detect patterns of fine phonetic detail in online speech perception. They differ only in the extent to which listeners

make syntactically informed predictions about whether to expect that pattern in upcoming speech.

1.5 Strategic listeners

A third perspective derives from the fact that cognitive capacity is limited. Monitoring, predicting, and processing incoming information across many different types of linguistic representation is cognitively expensive, and it is not obvious that listeners are always willing to expend that effort in all circumstances. This final view, which I shall call the *strategic listener* perspective, holds that listeners' attention to different types of information can change dynamically, adapting to circumstances. Kim et al. (2020), for example, demonstrated this adaptability in a perceptual study focusing on the difference between the vowels [ɛ] and [æ]. When the stimuli were manipulated to reduce spectral cues to vowel identity, listeners shifted their attention to durational cues. Intriguingly, exploratory analyses suggested that this flexible use of cues was larger for listeners with weaker inhibitory control. The authors propose that this pattern reflects a tendency for weak inhibitory control to be associated with the ability to keep attentional focus broad, across a larger variety of information. This breadth of attentional focus could have facilitated the shift in relative cue-weighting for listeners with low inhibitory control. Rather than disregarding less important cues to vowel identity, such as duration, they were more likely to attend to it, and so were well-placed to up-weight its importance when spectral cues became unreliable.

When cognitive resources are taxed, attention to acoustic detail is reduced. Mattys and colleagues showed that increasing cognitive load interferes with a listener's ability to process low-level acoustic detail, shifting focus instead to lexico-semantic information (Mattys & Wiget, 2011; Mattys et al., 2009, 2014). Data from Clayards et al. (2021) further suggests that this flexible attention to cues may be deployed strategically: Over the course of an experiment, listeners became less responsive to fine phonetic detail distinguishing the initial syllable in words like *discolor* from its counterpart in words like *discover*. The authors propose that this reduced use of phonetic cues reflects rapid, strategic adaptation to the task: Listeners learned that the rest of the sentence would disambiguate the target words, thus obviating the need to attend to phonetic detail so closely.

Christiansen & Chater (2016) describe the need to keep up with the ever-changing flood of input in spoken language processing as the 'Now-or-Never Bottleneck', which listeners navigate through strategic compression. One of the core principles of the Now-or-Never Bottleneck is that listeners process input according to a hierarchical organization of linguistic representations from lower-level, less abstract units, such as the phonetic speech stream, into higher-level, more abstract units, such as words and discourse. Incoming 'chunks' of perceptual information are compressed rapidly into abstract, higher-level units, and passed up to other levels of representation to free up low-level processing capacity for new input. Low-level information lost during compression

is lost for good: If it's not processed now, it is never processed at all. Under this view, dynamic changes in perceptual processing reflect changes in which information gets lost during the initial compression. When processing demands increase, but the real-time flow of new input does not abate, listeners must compress more of the low-level information to keep up, which means that more acoustic detail is lost during compression. It is this trade-off that is responsible for the observed patterns, such as reduced use of acoustic-phonetic detail under high cognitive load conditions (Mattys & Wiget, 2011; Mattys et al., 2009, 2014).

Kuperberg & Jaeger (2016) propose that the trade-off between cognitive capacity limitations and information processing might be understood in terms of a utility function. This utility function weighs the benefits of accurately predicting upcoming information from previous cues against the 'metabolic costs' associated with forming those predictions (p. 44). Thus, listeners may disregard certain types of cues that are not immediately relevant to the listening task at hand, such as semantic information in a phoneme monitoring task. They may further disregard or down-weight those cues which are not as reliable as other cues at hand, or carry less new information in a given context. The benefit of using these cues is not worth the effort of making the predictions. Importantly, optimizing these metabolic costs can influence the types of cues listeners attend to in real-time speech processing. Since acoustic input changes quickly, listeners making predictions about upcoming phonetic information must make many more predictions for a given stretch of speech than they would need to make if they were predicting syntactic structures. A rational listener, according to Kuperberg & Jaeger (2016), would therefore hesitate to predict upcoming phonetic information unless the benefits to perceptual processing are sufficient to justify the metabolic cost associated with making those predictions.

Consider now how strategic listeners would handle sentences like (1–2), repeated below for ease of reference.

(1) a. It seems that the rock attracts migrating songbirds.
b. It seems that the rocks attract migrating songbirds.

(2) a. It seems that this rock attracts migrating songbirds.
b. It seems that these rocks attract migrating songbirds.

In sentences like (1a–b), the lack of morphosyntactic information in the determiner means that listeners would need to wait until the very end of the noun to determine whether it was unsuffixed or suffixed, singular or plural. In such a context, they might well find it worth their while to attend to the rapidly-changing phonetic information in the stem. Even though they must use it to form an additional prediction quite rapidly, incurring an additional metabolic cost (Kuperberg & Jaeger, 2016), that prediction still gives them a head start on building their expectations about the grammatical number of the noun in question. By contrast, in sentences like (2a–b), the demonstrative already provides ample information about the number of the noun, and the suffix

or lack thereof will offer further confirmation. In these sentences, there is little benefit to using this phonetic information to make further predictions about the form or number of the noun. Unlike phonetic predictors, a strategic listener would not bother expending the effort to predict the lengthening associated with an unsuffixed singular stem, or the shortening associated with a suffixed plural stem. And unlike baseline ideal observers, they would not bother attending to that detail to form predictions about the likely presence or absence of the upcoming suffix. In sum, a strategic listener may be more attentive to segmental compression patterns after non-agreeing determiners – where they carry more information – than after agreeing determiners.

1.6 The current study

The current study was designed to test whether listeners use their knowledge of segmental compression to the same extent as polysyllabic shortening, and, if so, whether that use depends upon morphosyntactic context. The baseline ideal observer account predicts that listeners should use phonetic compression similarly, regardless of morphosyntactic context. The phonetic predictor account holds that listeners should use phonetic compression more after agreeing determiners than non-agreeing determiners; and the strategic listener account predicts that listeners should use phonetic compression more after non-agreeing determiners than after agreeing determiners.

To test these three accounts, I constructed a visual-world eye-tracking study, and asked listeners to click on a target picture whose name was embedded in sentences like (3–4). Note that the (a) sentences will serve as the critical stimuli, while the (b–c) sentences will serve as fillers.

(3) a. It seems that *this rock* attracts migrating songbirds.
b. It seems that *these rocks* attract migrating songbirds.
c. It seems that *the rocket* was set off by mistake.

(4) a. It seems that *the rock* attracts migrating songbirds.
b. It seems that *the rocks* attract migrating songbirds.
c. It seems that *the rocket* was set off by mistake.

In such sentences, lengthening the duration of a singular target stem in (3a) and (4a) should facilitate listeners in distinguishing it from plural and two-syllable competitors, because the longer duration matches listener expectations about one-syllable, unsuffixed noun stem duration. By contrast, shortening the duration of the stem should cause more difficulty in disambiguating a singular target from competitors, because the shortening should lead listeners to expect a plural suffix or further syllables – a prediction which does not match the eventual speech signal. This pattern of superior identification of singular targets when their stems are lengthened rather than shortened shall be called the *Match effect*. By comparing the Match effect when listeners must distinguish between singular targets and plural competitors to the Match effect when

they must distinguish between singular targets and two-syllable competitors, we will be able to compare listeners' use of segmental compression to polysyllabic shortening in their online speech perception. Then, by comparing the Match effect after agreeing determiners, as in (3), relative to non-agreeing determiners, as in (4), we will be able to distinguish between the baseline ideal observer, the phonetic predictor, and the strategic listener accounts.

2. Methods

2.1 Materials

Critical stimuli were built around a set of 84 stems, which appeared in one of three forms: a singular (e.g., *rock*), a plural (*rocks*), and a two-syllable word (*rocket*). The singular served as the target in critical trials, as in (3a) and (4a), while the plurals and two-syllable words were the targets of filler trials, as in (3b–c) and (4b–c). These triplets were selected to satisfy to the following four priorities:

- The singular had to end in a non-sibilant consonant, to avoid the */-əz/* allomorph of the suffix, which would increase the syllable count.
- All three members of the triplet needed to be nouns, given the use of the determiner to manipulate syntactic context.
- All three members of the triplet needed to be sufficiently imageable for use in a visual-world eye-tracking paradigm.
- The two-syllable word needed to lexically embed the singular stem in its first syllable, *without* using morphological derivation to achieve the embedding. This allowed the materials to fully dissociate the morphological distinction between singulars and plurals from the syllable-count distinction between singulars and two-syllable words.[3]

Given these constraints, the 84 stems selected represented approximately the full range of such triplets available in the English vocabulary.[4] For this reason, it was not possible to further control for properties such as frequency or prosodic structure.

In Agreeing contexts, each target word appeared after an agreeing determiner (e.g., *this/that* for singular and two-syllable targets, and *these/those* for plural targets), while in Non-Agreeing contexts, they were uniformly preceded by *the*. In all sentences, the word immediately following the singular target began with the same initial phonemes as the continuation of the two-syllable word. For example, the portion of the sentence following the singular target *rock* in both the

[3] This condition was not met by *lock/locker*.

[4] This limit is accent-dependent. Accents of English with the *merry/marry/Mary* merger would allow pairs such as *bear/berry* and *chair/cherry*. Accents that do not palatalize /t/ before /ju/ (or lack /ju/ altogether) would allow *tool/tulip*.

Agreeing sentence frame (3a) and the Non-Agreeing sentence frame (4a) begins with *[ət]*. This preserves a segmental ambiguity between the competitor *rocket* ([ˈɹɑkət]) and the sequence in the stimulus sentence *rock attracts* ([ˈɹɑkəˈʧɹakts]).[5] In this particular example, the following context fully matched the second syllable of the two-syllable word, but in most cases it was not possible to match more than the onset and nucleus of the second syllable, or occasionally just the initial onset. Forcing a more complete overlap in the following phonological environment would have required torturing the sentences beyond any hope of producing naturalistic stimuli. A full list of sentences is provided in the Appendix, in Table 7.

Matching the following phonological context with the second syllable of the two-syllable word served to minimize any coarticulation bleeding into the noun stem from the following word. This removed a disambiguating cue that listeners might otherwise use to exclude the two-syllable competitor from consideration. It also served to maximize the period in which listeners would have to draw solely on durational cues to identify the target before later phonemic information became available. One consequence of this manipulation is that singulars were disambiguated from plural competitors earlier, at the end of the stem, than from two-syllable competitors. I return to this point in the Section 4.

The raw recordings, therefore, comprised a set of six sentences for each stem, crossing the three word forms (singular, plural, two-syllable) with the two sentence contexts (Agreeing, Non-Agreeing) to produce a set of the sort shown in (3–4). All recordings were produced in a sound-attenuated booth by a native male speaker of Scottish Standard English, who was not aware of the experimental focus on polysyllabic shortening and segmental compression. He was instructed to produce all sentences in such a way as to make the six versions of each item as similar as possible. He read each set of sentences twice, and the recordings with the most similar prosodic structure across the three word types were selected for further analysis and stimulus-creation.

2.2 Acoustic analysis

Since Scottish Standard English has not previously been used in studies of polysyllabic shortening and segmental compression, an initial analysis of the selected raw recordings was conducted to confirm that the speaker's natural, unmanipulated speech still contained the durational patterns of interest. Accordingly, the raw stem durations (see **Table 1**) were extracted, log-transformed, and analyzed with linear mixed effects regression in Julia (version 1.10.2; Bezanson et al., 2017), using the package `MixedModels` (version 4.24.0; Bates et al., 2024).

[5] Although the /t/ in *attracts* does become palatalized to [ʧ] before [ɹ], its onset is sufficiently similar to [t] to ensure that the segmental ambiguity between *rocket* and *rock attracts* extends beyond the offset of the target word *rock*.

Table 1: Mean stem durations (and standard deviations) in ms from the raw recordings.

Sentence frame	**Voiceless stems**			**Voiced stems**		
	singular	**plural**	**two-syllable**	**singular**	**plural**	**two-syllable**
context	253 (55)	244 (43)	228 (43)	305 (46)	262 (44)	295 (49)
dur. dif. from sg.	0 (0)	8 (35)	25 (41)	0 (0)	43 (43)	11 (33)
no context	300 (57)	262 (46)	249 (42)	337 (55)	281 (42)	312 (45)
dur. dif. from sg.	0 (0)	37 (33)	50 (41)	0 (0)	56 (38)	26 (37)

Independent variables included Word Type, Context, and Voicing. The values for Word Type were Singular, Plural, and Two-Syllable, as described in 2.1. The values for Context were Agreeing, which coded sentences containing agreeing demonstrative determiners, and Non-Agreeing, which coded sentences using the non-agreeing determiner *the*. Finally, Voicing coded for whether the final obstruent of the stem was Voiced or Voiceless, to account for the well-known fact that the voicing of a coda consonant or consonant cluster produces substantial, highly perceptible durational effects on the preceding vowel (Raphael, 1972).

All variables were contrast-coded, to ensure that coefficient estimates for simple effects reflected proper main effects (Brehm & Alday, 2022). Context and Voicing were effects-coded, with Agreeing and Voiced set at 1, and Non-Agreeing and Voiceless set at -1. Thus, positive model coefficients indicate longer stem duration for Context compared to No-Context, and for Voiced compared to Voiceless, while negative model coefficients indicate the reverse. Word-type was Helmert-coded. The first contrast compared Two-Syllable stem duration with the combination of Singular and Plural stem duration, to confirm that the polysyllabic shortening pattern was present. The second contrast compared Plural stem duration with Singular stem duration, to confirm that the expected segmental compression pattern was present.

The model's fixed effects included all three variables, plus their higher-order interactions, while the random effects included intercepts by stem, along with random slopes for the interaction Word Type and Context. As Voicing did not vary within stems, it was not included as a random slope.

The final model summary is provided in **Table 2**, with its effects visualized in **Figure 1**. The model revealed that, as expected, Two-Syllable durations were shorter than Plural and Singular durations ($\beta = -0.038$, $p = .004$), reflecting a polysyllabic shortening pattern. Plural durations were also quite substantially shorter than Singular durations ($\beta = -0.123$, $p < .001$), reflecting a robust segmental compression pattern – which, as **Figure 1** shows, seemed to be even stronger than the polysyllabic shortening pattern for stems with voiced final sounds. Whereas voiceless stem durations (right panel, **Figure 1**) were shortest in Two-Syllable words, voiced

stem durations (left panel) were shortest in Plurals. The negative coefficient for Context indicates that stems were shorter after Agreeing determiners than after Non-Agreeing *the* ($\beta = -0.048$, $p < .001$), while the positive coefficient for Voicing confirms that, as expected, stems were longer when they ended in voiced sounds than voiceless obstruents ($\beta = 0.081$, $p < .001$).

Table 2: Summary of regression model of log-transformed stem durations in the raw recordings.

	Dependent variable: log duration			
Variable	**Estimate**	**(SE)**	***z***	***p***
Intercept	−1.30	(0.017)	−75.95	<.001
Word Type (2syll.vs.others)	−0.038	(0.013)	−2.86	.004
Word Type (pl.vs.sg)	−0.123	(0.011)	−11.18	<.001
Context	−0.048	(0.005)	−9.92	<.001
Voicing	0.081	(0.017)	4.74	<.001
Word Type (2syll.vs.others) × Context	0.014	(0.007)	1.85	.064
Word Type (pl.vs.sg) × Context	0.032	(0.009)	3.56	<.001
Word Type (2syll.vs.others) × Voicing	0.064	(0.013)	4.82	<.001
Word Type (pl.vs.sg) × Voicing	−0.045	(0.011)	−4.05	<.001
Context × Voicing	0.009	(0.005)	1.97	.049
Word Type (2syl.vs.others) × Context × Voicing	−0.002	(0.007)	−0.21	.833
Word Type (pl.vs.sg) × Context × Voicing	−0.02	0.009	−2.16	.031
Observations		546		
Log Likelihood		379.849		
Akaike Inf. Crit.		−691.698		
Bayesian Inf. Crit.		−545.409		

The interactions are visualized in **Figure 1**. The positive interaction terms for Word Type by Context indicate that the effect of Word Type was larger for Non-Agreeing than Agreeing conditions. This was a non-significant tendency for the difference between Two-Syllable against other Word Types ($\beta = 0.014$, $p = .064$), but significant for the difference between Plural and Singular ($\beta = 0.032$, $p < .001$). The positive interaction term between Voicing and the contrast between Two-Syllable against other Word Types corresponds to a stronger polysyllabic shortening effect with Voiceless stems compared to Voiced stems ($\beta = 0.127$, $p < .001$). The

negative interaction term for Plural against Singular indicates that the segmental compression effect was weaker in Voiceless stems relative to Voiced stems ($\beta = -0.089$, $p < .001$). The positive interaction term between Context and Voicing suggests that the effect of Context may be slightly larger with Voiceless stems than with Voiced stems. Finally, the negative three-way interaction ($\beta = -0.02$, $p = .031$) reveals that the difference between Plural and Singular stems, which is exaggerated in Non-Agreeing compared to Agreeing contexts, is not quite so exaggerated with Voiced stems as it is with Voiceless stems.

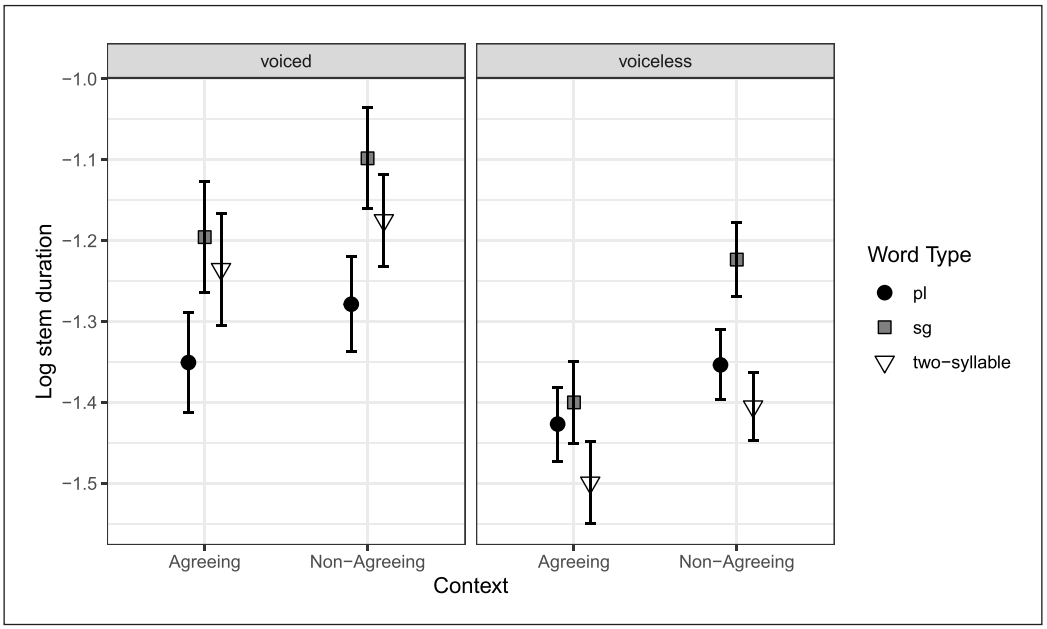

Figure 1: Model effects plot for the interactions between Word Type, Context and Voicing in the recorded stimuli. Error bars indicate 95% confidence intervals.

With this confirmation in hand that the raw speech of the speaker contains both the polysyllabic shortening and segmental compression patterns of interest in its natural form – albeit in varying degrees across Context and Voicing – the next step was to manipulate the speech to produce the desired experimental conditions.

2.3 Acoustic manipulation

For each item, each of the six raw sentences was segmented in Praat (Boersma & Weenink, 2015) into the following component elements: The preamble consisted of everything up to the onset of the determiner. The determiner followed, with the stem next, followed by the suffix in plural sentences and the rest of the word in sentences containing two-syllable targets. The postamble consisted of everything after the offset of the word (**Figure 2**, step 1). The stem was further

subdivided into onsets, vowels, sonorant codas, and obstruent codas. Segmentation criteria generally followed recommendations from Turk et al. (2006), and are described in more detail in the supplemental materials on the Open Science Framework (OSF) archive for this project.

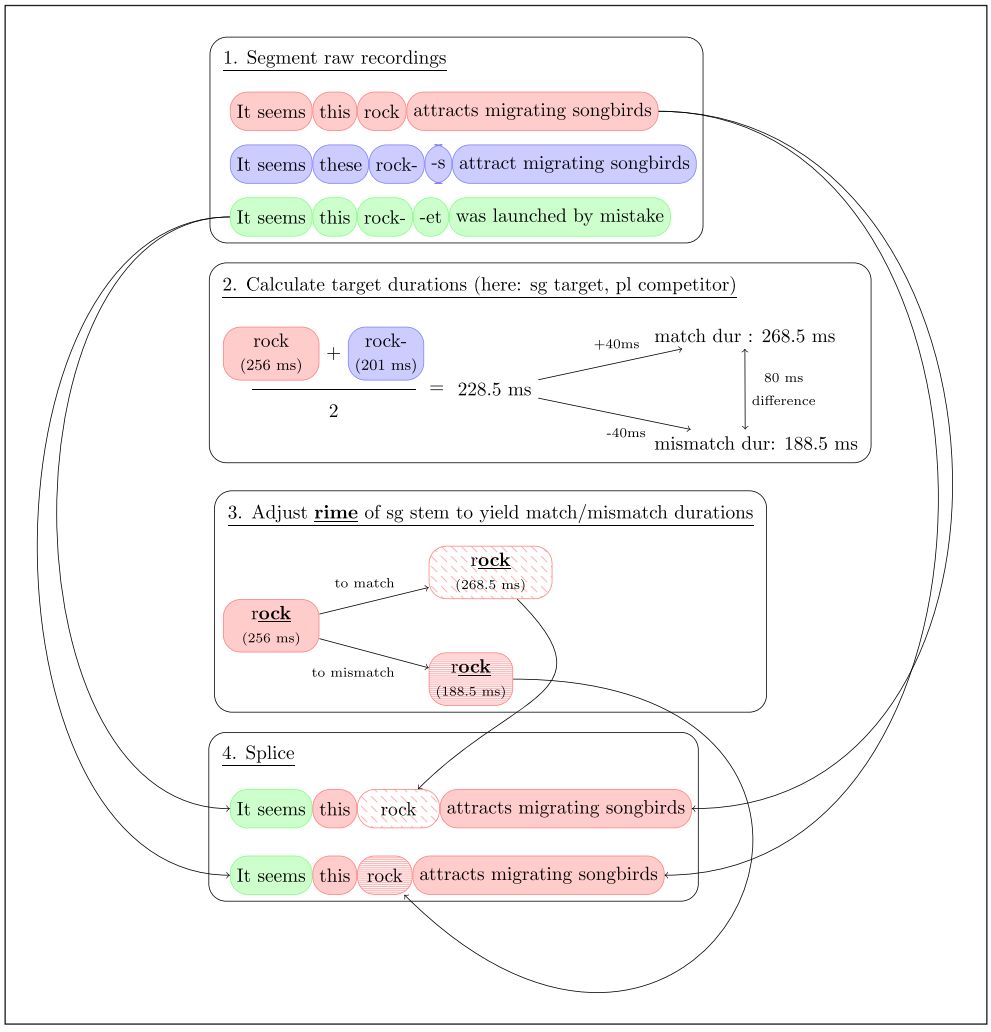

Figure 2: Sample algorithm for creating critical stimuli containing singular targets designed to be presented with a plural competitor. (1) Raw sentences are first segmented into preambles, determiners, stems, and postambles. (2) The target durations for Match and Mismatch stems are then calculated in order to achieve an 80 ms difference between the Match and Mismatch stems, with the Match stems 40 ms longer than the average duration of the singular and plural raw durations, and the Mismatch stems 40 ms shorter. (3) The duration of the rime of the extracted singular stem is adjusted so that the full stem yields the target Match and Mismatch durations. (4) The sentence components are spliced together, forming Match and Mismatch stimuli. In this case, the preamble is used from sentences recorded with two-syllable targets, followed by the original determiner from the singular raw recording, followed by the adjusted stem, and finishing up with the original postamble from the singular raw recording.

Next, the duration of each critical target – which was always the singular stem – had to be set. Within each sentence, the critical singular target needed to be adjusted to appear in a Match or a Mismatch condition. In the Match condition, the singular target was lengthened, so that it contained duration cues consistent with a word boundary immediately following the stem. In other words, Match targets were designed to signal that the word was singular. In the Mismatch condition, the stem was shortened, so that its acoustic cues signaled that something would follow the stem – either an additional segment on the syllable coda, such as a plural suffix, or else an additional full syllable. In other words, Mismatch targets were designed to mislead the listener into expecting a plural or two-syllable noun.

The raw durational differences between singular and two-syllable stems were not consistently higher or lower than the raw durational differences between singular and plural stems. When stems were voiceless, two-syllable stems were shorter than singulars to a greater degree than plural stems, while when stems were voiced, it was plural stems that showed the greatest degree of shortening. For this reason, the same durational manipulation was used to distinguish Match and Mismatch targets, regardless of the nature of the competitor object. I settled on an 80 ms difference, which was large enough to be detectable, but not so extreme as to sound unnatural to my ear. Previous pilot work had determined that Match effects were elicited by durational differences of 60 ms and 90 ms, so the final difference of 80 ms used here fell within that range.

The procedure for manipulating the duration of the rimes is visualized in **Figure 2**, steps 2–3. This was done twice for each stem, depending on the intended competitor type. In a trial with a singular target and a plural competitor, the raw durations of the singular and plural recordings were averaged together; to this was added 40 ms to arrive at the Match duration for the singular, and from this was subtracted 40 ms to arrive at the Mismatch duration for the singular. In a trial with a singular target and two-syllable competitor, the singular and two-syllable raw stem durations were averaged together, after which everything proceeded in the same way as before (**Figure 2**, steps 2–3).

The durations of all stem rimes were then adjusted to make the stem match the target Match or Mismatch duration, using the overlap-add implementation of PSOLA in Praat. I applied the durational manipulation only to the rime of the singular stem – i.e., vowel + coda, treated as a single unit – because previous work on polysyllabic shortening patterns in Dutch has suggested that the bulk of the durational effects are carried by the rime, while the onset is invariant across forms (Kemps, Ernestus, et al., 2005). This invariant onset possibly serves to provide listeners with a baseline against which the rest of the stem durational patterns can be evaluated. English is not, of course, Dutch, but the two languages show a great deal of similarity in listeners' use of polysyllabic shortening patterns (for Dutch, see Salverda et al., 2003; Shatzman & McQueen, 2006; for English, Blazej & Cohen-Goldberg, 2015; Davis et al., 2002; for both together, see Kemps, Wurm, et al., 2005). More practically, adjusting the duration of the entire stem resulted

in some unnatural-sounding tokens, particularly those beginning with voiceless stops, while limiting the duration to the rime alone produced better-sounding results.

Finally, it was necessary to ensure that the preamble contained no cues biasing the listener to expect either the singular target (*rock*) or its competitor. If the competitor was plural, then neither the preamble from the raw singular recording (as in (3a) or (4a)) nor the preamble from the raw plural recording ((3b) or (4b)) could be used. Therefore, for singular targets with plural competitors, the final sentence was constructed by splicing the determiner, manipulated stem and following postamble onto the preamble of the raw two-syllable recording ((3c or 4c)). This ensured that any cues in the preamble to the identity of the upcoming noun would point to the two-syllable word (*rocket*), which is not present on the screen, rather than to the target (*rock*) or the competitor (*rocks*). By the same logic, when the competitor was a two-syllable word (*rocket*), the determiner, manipulated stem, and postamble were spliced onto the preamble from the plural recording ((3b or 4b)), so that any unintentional cues in the preamble would point to a plural noun, absent from the screen, rather than a singular target or two-syllable competitor (**Figure 2**, step 4). This splicing also had the effect of ensuring that all possible speech rate cues were fully controlled, as the same preamble was used across different Match conditions. In this way, speech rate has no opportunity to affect how listeners process durational cues (Dilley & Pitt, 2010; Pitt et al., 2016).

Selections of these manipulated recordings were evaluated by members of the Glasgow University Laboratory of Phonetics at a lab meeting, and a full experimental list was further evaluated by a trained phonetician in the context of an experimental run. Both sets of evaluators were fully informed about the logic and design of the manipulations, and the lab group was further informed about the identity of each recording – whether it had been lengthened or shortened. The lab group were instructed to attend to the length contrast between Match and Mismatch conditions, while the trained phonetician was instructed to listen carefully for acoustic artifacts as the experiment unfolded. Both groups reported that the manipulations in question (durational contrast for the lab group; acoustic artifacts for the trained phonetician) sounded natural. Experimental participants were all asked about the quality of the audio, and although they volunteered comments about the structure of the sentences, the soothing quality of the speaker's voice, and their subjective experience of being temporarily confused between singular and two-syllable stems, none reported noticing a durational manipulation, and all agreed that the quality of the audio was excellent.

Filler sentences – those containing plural or two-syllable targets – were created according to the same procedure, differing only in that the stems for the plural and two-syllable fillers appeared only in the shortened form. This ensured that the noun target in all sentences was manipulated, avoiding any systematic auditory difference between the critical and filler targets. Further, this manipulation provided a crucial experimental control. As Clayards et al. (2021)

observed, listeners might reduce their use of durational cues across the course of an experiment if they learn that those cues are unreliable. No matter how uniformly, predictively, or strategically listeners can deploy their knowledge of polysyllabic shortening and segmental compression patterns, they may cease to do so if they repeatedly encounter critical stimuli whose Match manipulations render such patterns useless. Presenting all filler targets with a shortened stem mitigates this problem, because it reinforces the fundamental experimental assumption that there is a roughly 80 ms difference between a singular stem and a plural or two-syllable stem. In filler trials, with plural or two-syllable targets, this pattern was uniformly true. In critical trials, with singular targets, this pattern was true half the time. With a 50/50 split between critical and filler trials, the entire experimental list, therefore, tested participants' ability to make use of a pattern that was 75% reliable. Thus, whatever listeners do with polysyllabic shortening and segmental compression, they are unlikely to stop doing it on the basis of the durational manipulation in the critical stimuli.

Plural fillers were spliced onto two-syllable preambles, parallel to their singular counterparts, and two-syllable fillers were spliced onto plural preambles. Thus, not only were all speech rate cues controlled across Match and Mismatch critical stimuli, they were also controlled across critical and filler stimuli.

2.4 Design

For each of the 84 stems, twelve critical sentences were created. Four of these contained a singular target whose duration was manipulated against its two-syllable competitor. These four sentences represented four conditions, formed by crossing the factors of sentence Context (Agreeing/Non-Agreeing) with duration Match (Match/Mismatch). Another four sentences all contained a singular target whose duration was manipulated against the plural competitor. They likewise represented the four conditions formed by crossing sentence Context and duration Match. These eight sentences comprised the audio stimuli for each item in a critical trial.

The final four sentences constructed for each stem formed the audio stimuli for filler trials. These contained the plural or two-syllable word as the target, and they varied across the two Context conditions. However, they only contained the Match duration, so that filler trials served to reinforce the expected polysyllabic shortening or segmental compression patterns. Each experimental list contained 42 critical and 42 filler trials. Stems were rotated across experimental lists in a Latin Square design, to ensure that all stems appeared in all conditions for critical trials, while each participant saw each stem only once.

The visual world for each trial contained four images: the target, its competitor, and two distractors. On critical trials, the target was always singular, and its competitor could be either plural or two-syllable. On filler trials, the target was either plural or two-syllable, and its

competitor could be singular, two-syllable (if the target was plural), or plural (if the target was two-syllable).

Distractor pictures were selected so that in the critical trials, the relationship between the target and competitor was mirrored in the distractors. Thus, if a critical trial had a singular target and plural competitor, the distractor pictures also contained a singular and plural version of another stem. If a critical trial had a singular target and a two-syllable competitor, the distractor pictures were themselves a singular word and its two-syllable counterpart. In filler trials, the relationship between target and competitor was mirrored in the distractors half the time, but not the other half of the time. These fillers ensured that participants would not be able to reliably recognize that some trials were about singular vs. plural, while others were about singular vs. two-syllable, because a full quarter of all trials mixed word types.

For the rest of this article, singular targets with plural competitors will be called *num-targets*, as the key feature distinguishing target and competitor is grammatical number. Singular targets with two-syllable competitors will be called *syll-targets*, because the feature distinguishing these targets from their competitors is syllable count.

2.5 Participants

The experimental protocol was approved by the College of Arts Ethics Committee at the University of Glasgow. Fifty-one participants (28 female, 18 male, 3 other; mean age = 27, ranging from 18 to 67) were recruited from the University of Glasgow community, as well as through Glasgow-specific social media advertisements (specifically, the r/Glasgow and r/GlasgowMarket subreddits). Data from two subjects was discarded, one for not being a native Scottish English speaker, and another for a voluntarily disclosed early-onset pseudo-dementia diagnosis. All remaining participants were native Scottish English speakers, with normal or corrected-to-normal vision and hearing. Participants were compensated with £12 for their time.

2.6 Procedure

The experimental procedure began with the collection of informed consent from participants, after which they filled out a brief language history questionnaire, and completed the visual-world eye-tracking experiment.

Eye-tracking data was collected from the right eye at a 1000 Hz sampling rate with an Eyelink 1000+ camera; stimuli were presented using the Experiment Builder software, with audio stimuli playing binaurally over headphones at a comfortable volume. Participants were informed that they would see four images appear on the screen, and hear a sentence mentioning one of the images. They were instructed to use the mouse to select the picture on the screen that was named in the sentence. Each experiment began with a nine-point calibration and validation,

which was repeated at the beginning of each block. The experiment began with seven practice trials before the 84 trials of the experiment itself, which were split into three blocks of 25, 25, and 34. Fillers and critical trials were randomly presented in each list, with the restriction that no particular combination of target type (singular, plural, two-syllable) and competitor type was presented more than twice in a row.

Each trial began with a drift correction fixation point on the screen. After participants fixated upon the point, four images appeared on the screen, each labeled underneath the image. Participants had 3.5 seconds to look freely at this screen to familiarize themselves with the pictures' names and locations on the screen before the audio began to play. Participants could not select an image before the audio began to play, but they were free to click on an image before the audio was completed. If they selected the correct image, a green frame appeared around the selected image; if they selected the wrong image, a red frame appeared around the selected image. Regardless of their selection, the entire audio file played to completion before the screen reset to the drift correction fixation point.

2.7 Analysis

2.7.1 Reaction time and accuracy

Reaction time and accuracy in clicking the target option were analyzed with (generalized) linear mixed models in Julia (version 1.10.2; Bezanson et al., 2017), as implemented in the `MixedModels` package (version 4.24.0; Bates et al., 2024).

Accuracy was analyzed with mixed effects logistic regression, with fixed effects of Context (Agreeing/Non-Agreeing), Match (Match/Mismatch), and Target Type (Syll-target/Num-target), along with all higher-order interactions. Random effects included random intercepts by subjects and stem, along with random slopes for all three variables and their higher-order interactions. Match, Context, and Target Type were all effects-coded. For Match, Mismatch was set to -1 and Match was set to 1; for Context, Non-Agreeing was set to -1 and Agreeing to 1; and for Target Type, Syll-target was set to -1 and Num-target to 1.

Reaction times were measured from the onset of the target noun. All correct RT measurements were then log-transformed and analyzed with linear mixed effects regression modeling, with the same coding and specification details as the accuracy analysis.

The model specifications used the maximal structure in order to test for the predicted three-way interaction that would fall out from the phonetic predictor and strategic listener accounts. The phonetic predictor account predicts a larger Match effect after agreeing determiners than non-agreeing determiners, but only for num-targets. The strategic listener account goes in the other direction: It predicts a larger Match effect after *non*-agreeing determiners than after agreeing determiners, but, again, only for num-targets.

2.7.2 Gaze traces

To analyze the gaze data, participant gazes were first binned into time windows. Since the duration of the determiners varied, with *the* being substantially shorter than demonstrative determiners, time windows during the determiner were simply set to cover the first half and second half of the determiner duration. Thus, the offset of the determiner occurred at the end of time window 2, regardless of the actual duration. Target word durations also varied across items, and within items, too, depending on whether they were in the Match or Mismatch condition. Therefore, time windows during the target word were also defined proportionally – this time as quintiles of the word's duration. Thus, the third time window covered the first fifth of the target word's duration, regardless of whether it was a Match or Mismatch word, the fourth time window covered the second fifth, and so on. The end of the seventh time window thus corresponded to the offset of the target word.

Time windows after the offset of the target were then set at 50 ms, to reduce the inherent dependencies holding between gaze positions separated by a single millisecond (Barr, 2008). Visual inspection of the overall averaged gaze trajectories revealed that the proportion of looks to syll-targets reached their peak by the 30th time window, or about 1100 ms after target offset, and for num-targets by the 25th window, or 850 ms after target offset. Accordingly, only the first 30 time windows were retained for statistical analysis for syll-targets, and only the first 25 time windows were retained for num-targets.

Within each time window, the proportion of looks to target and looks to competitor was calculated by taking the number of samples in which the gaze position rested on the target or competitor interest area, and dividing by the number of readable samples within the time window. This calculation had the effect of removing blinks or other unreadable samples. Samples registered during saccades were retained, as many saccades moved from one location to another within the same interest area – e.g., shifting from the label at the bottom of the image to the image itself. All proportions were then converted into empirical logits (elogs) (Barr, 2008). Finally, the target advantage was calculated by subtracting the elog-converted looks to competitor from the elog-converted looks to the target. This difference – henceforth termed *target advantage* – provided the dependent variable in the gaze-trace analysis.

Even with the time-window binning strategy, the distribution of the dependent variable was far from normal. Instead, it had a trimodal distribution, with peaks near -10, 10, and 0 (**Figure 3**). These peaks correspond to a preponderance of time windows in which listeners are looking, respectively, only at the competitor, only at the target, or at something else entirely, such as a distractor or the central fixation cross. Only time windows containing saccades with either target or competitor as the endpoint would produce values outside these three peaks. Yet regression models which estimate means of the dependent variable assume the normality of the dependent variable; and that assumption is not met here. Therefore, the data were analyzed with empirical

Bayesian non-parametric quantile regression modeling (QGAMs; Baayen et al., 2022; Fasiolo et al., 2021). QGAMs are an extension of generalized additive mixed modeling methods (GAMMs; Sóskuthy, 2017; Wood, 2011, 2017), with the advantage that they do not rely on assumptions about the distribution of the dependent variable.

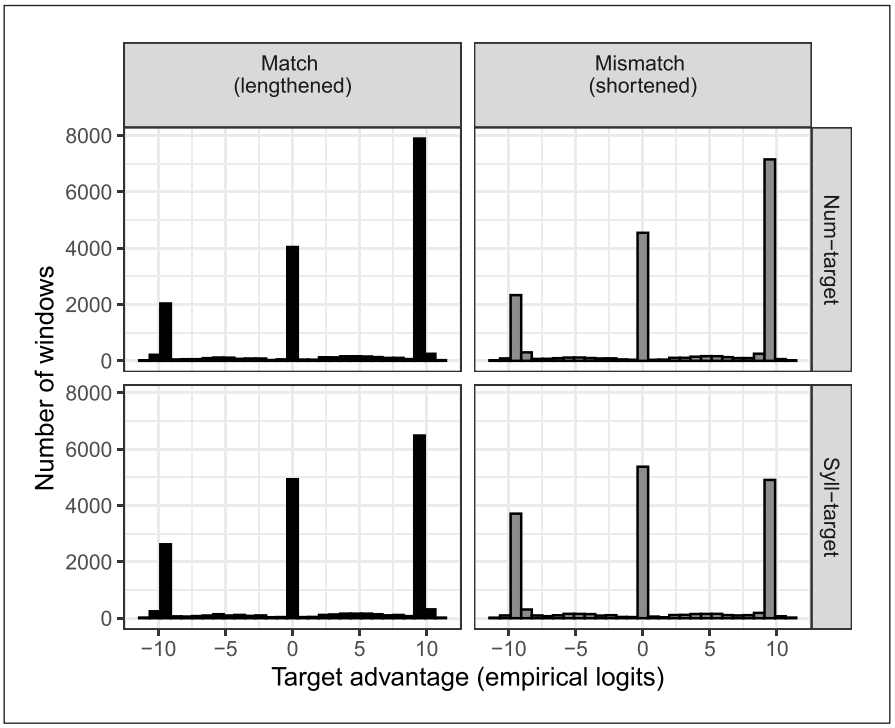

Figure 3: Histogram of target advantage values in empirical logits in the binned gaze data, separated by target type (Num-targets on top; Syll-targets on bottom) and Match (lengthened Match targets on left, shortened Mismatch targets on right). All binned gaze data is profoundly trimodal, with two peaks at either end of the distribution, and one at 0.

QGAMs estimate not mean values of the dependent variable, but a given quantile. Here, I report the results for the median, although the supplemental materials on the OSF archive for this project provide additional analyses examining the effects of Match and Context across different quantiles of target advantage. All analyses were conducted in the `qgam` package (version 1.3.4; Fasiolo et al., 2021). Syll-targets were analyzed with separate models from num-targets, but the analysis structure for each target type was comparable. In each QGAM, two types of predictor variables were included: parametric terms and smooth terms. The parametric terms operate similarly to terms in a linear model. Their coefficients capture the overall (median) height of the curve across the entire time period of interest (here, the first 25 or 30 time windows). Context and Match were both effects-coded, with the values Agreeing and Match set at 0.5, and Non-Agreeing and Mismatch set at -0.5. Thus, positive parametric coefficients for Context indicate

higher overall values of target advantage for Agreeing sentences (in which the stem is preceded by *this/that*) than for Non-Agreeing sentences (in which the stem is preceded by *the*), and positive parametric coefficients for Match indicate higher overall target advantage for lengthened Match stimuli than for shortened Mismatch stimuli.

The smooth terms capture the dynamic changes of gaze trajectories across that time period, allowing us to observe differences in the curve shapes – e.g., steeper or shallower increases that begin or peak at different time points – even if there is no overall difference in the gaze proportions through the period of interest. Thus, QGAMs not only allow us to avoid violating assumptions about the distribution of the dependent variable; they also allow us to understand the nature of the gaze traces in a more nuanced way than traditional ANOVAs, and are more flexible in capturing different types of curve shapes than the strictly polynomial terms used in growth-curve analysis (Mirman, 2014).

The smooth terms for key variables were specified with thin plate regression splines. Both models further included random factor smooths by subject, to control for individual variation across gaze trajectories, and random factor smooths by item, to control for item-specific effects springing from variation across prosody, phoneme overlap, lexical frequency, and so on.

Because QGAMs do not allow for the straightforward interaction between factors in the smooth terms, interactions in the smooths were examined by crossing the Context and Match terms to create a four-level factor. All factors in the smooth terms were treatment-coded. Because the three accounts described in 1.3 through 1.5 differ specifically in their predictions regarding the interaction between Match and Context, both models contained Match, Context, and their interactions in both parametric and smooth terms.

For num-targets, the baseline ideal observer account predicts no interaction between Match and Context; the phonetic predictor account predicts larger effects of Match in Agreeing conditions than Non-Agreeing conditions; and the strategic listener account predicts a larger effect of Match in Non-Agreeing conditions than Agreeing conditions. All three accounts, however, concur that there should be no interaction between Match and Context for syll-targets, because the presence or absence of a determiner like *that* is irrelevant to distinguishing a word like *cart* from a word like *carton*.

3. Results

3.1 Behavioral results

3.1.1 Accuracy

The model summary for the accuracy analysis is provided in **Table 3**, and a visualization of the partial effects is provided in **Figure 4**.

Table 3: Summary of the logistic regression model of accuracy. All variables are effects coded, with Mismatch, Non-Agreeing Context, and Syll-target Target Type set to −1, while Match, Agreeing Context, and Num-target Target Types set to 1. Positive coefficients reflect higher accuracy for Match, Agreeing Context, and Num-targets.

Dependent variable: accuracy				
Variable	**Estimate**	**(SE)**	***z***	***p***
Intercept	5.187	(0.506)	10.26	<.001
Context	−0.065	(0.208)	−0.31	.753
Target Type	2.34	(0.360)	6.49	<.001
Match	−0.203	(0.231)	−0.88	.380
Context × Target Type	0.130	(0.219)	0.60	.552
Match × Context	−0.109	(0.268)	−0.40	.686
Match × Target Type	−0.465	(0.236)	−1.97	.048
Match × Context × Target Type	0.024	(0.257)	0.09	.925
Observations	2,058			
Log Likelihood	−452.011			
Akaike Inf. Crit.	1064.021			
Bayesian Inf. Crit.	1514.380			

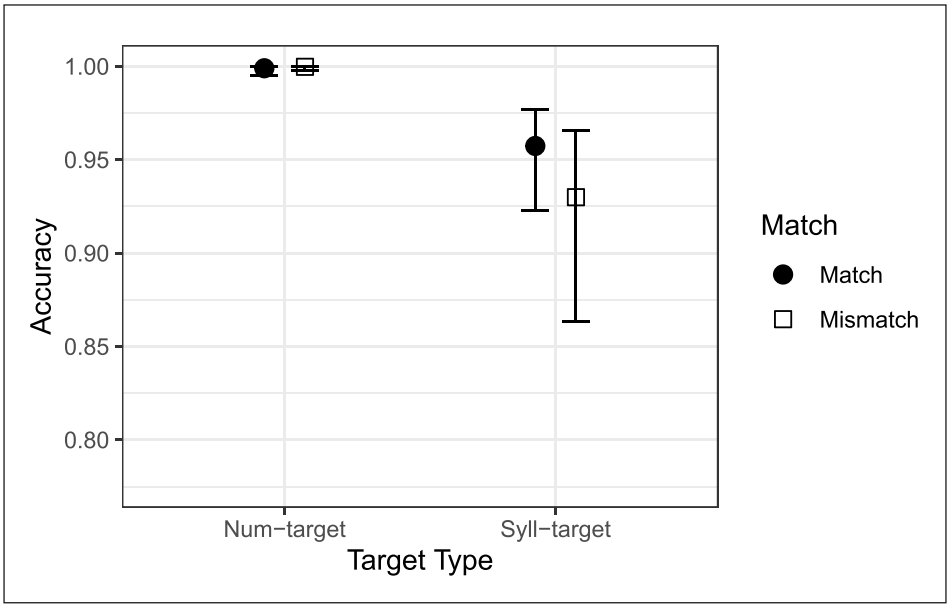

Figure 4: Partial effects plot for the accuracy model, showing interactions between Match and Target Type. Error bars indicate 95% confidence intervals.

The effect of Target Type reflected higher accuracy with Num-targets compared to Syll-targets ($\beta = 2.34$, $p < .001$). There was no significant main effect of Match, but the interaction between Match and Target Type reveals that Syll-targets did have higher accuracy with Match than with Mismatch targets ($\beta = -0.465$, $p = .048$). As **Figure 4** illustrates, it is unsurprising that Num-targets showed little effect of Match, as their accuracy was already at ceiling.

3.1.2 Reaction time

The model summary for the reaction time analysis is provided in **Table 4**, and a visualization of the partial effects is provided in **Figure 5**.

Table 4: Summary of the linear regression model of reaction time. All variables are effects coded, with Mismatch, Non-Agreeing Context, and Syll-target Target Type set to -1, while Match, Agreeing Context, and Num-target Target Types set to 1. Positive coefficients reflect slower RT accuracy for Match, Agreeing Context, and Num-targets.

Dependent variable: log reaction time				
Variable	**Estimate**	**(SE)**	***z***	***p***
Intercept	7.312	(0.037)	197.47	<.001
Context	−0.001	(0.007)	−0.13	.898
Target Type	−0.083	(0.009)	−9.23	<.001
Match	−0.032	(0.007)	−4.53	<.001
Context × Target Type	0.018	(0.0068)	−2.66	.008
Match × Context	−0.012	(0.008)	−1.58	0.114
Match × Target Type	0.014	(0.006)	2.15	0.021
Match × Context × Target Type	−0.002	(0.007)	−0.40	.689
Observations	1,820			
Log Likelihood	−198.067			
Akaike Inf. Crit.	558.139			
Bayesian Inf. Crit.	1004.173			

The effect of Target Type revealed faster RTs to Num-targets than Syll-targets, ($\beta = -0.083$, $p < .001$), while an effect of Match revealed faster RTs to Match targets relative to Mismatch targets ($\beta = –0.032$, $p < .001$). These effects were qualified by two interactions. The first, between Target Type and Context ($\beta = .018$, $p = .008$), reflects a larger effect of Target Type in Agreeing contexts (filled triangles, left side of **Figure 5**) compared to Non-Agreeing contexts (grey diamonds). The second interaction, between Match and Target Type ($\beta = .014$, $p = .021$),

reflects a larger Match effect for Syll-targets compared to Num-targets (right-hand panel of **Figure 5**).

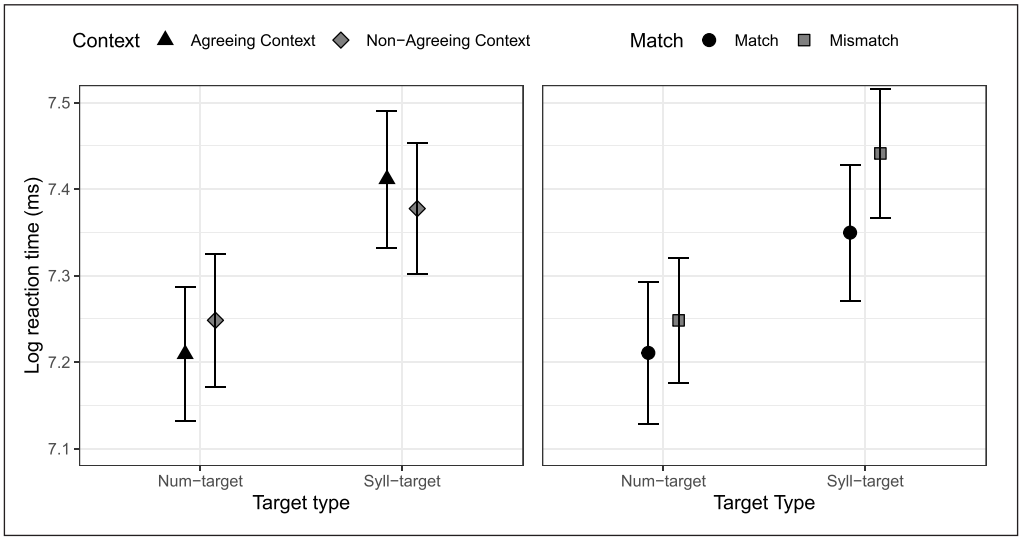

Figure 5: Partial effects plot for RT model, showing the interactions between Context and Target Type (right), and Match and Target Type (left). Error bars indicate 95% confidence intervals.

In summary, then, the behavioural results reveal higher accuracy and faster RTs for num-targets compared to syll-targets, and a larger Match effect for syll-targets compared to num-targets. There is also evidence from the interaction between Context and Target Type that participants are alert to the agreement information, as the presence of an agreeing determiner facilitates RTs to num-targets more than to syll-targets. However, there was no three-way interaction in the behavioural data to support the prediction that sensitivity to agreement modulates the Match effect, as predicted by the strategic listener and phonetic predictor accounts.

3.2 Gaze traces

3.2.1 Syll-targets

Table 5 shows the model summary of the analysis for syll-targets, which is visualized in the right-hand side of **Figure 6**.[6] The parametric effect of Match reveals that participants looked

[6] **Figure 6** presents only the model summary curves, without overlaying the raw data. This is for two reasons. First, the trimodal distribution of the data means that plotting median points from the raw data results in a visually unsatisfying set of clusters around −10, 0, and 10, rather than a smooth curve that tracks the model estimates. Plotting the means from the raw data produces a smooth curve that shows exactly the same pattern as the model results. However, the leftward skew of the distributions of target advantage, visible in **Figure 3**, means that the medians estimated from the model are higher than the means in the raw data in most combinations of conditions. The exception to this is

more towards the target in the Match condition than in the Mismatch condition ($\beta = 1.812$, $p < .001$). This can be observed in the right-hand panels of **Figure 6**, which show a clear target advantage for the gaze traces in the lengthened Match condition (dark ribbons) relative to the shortened Mismatch conditions (gray ribbons). The parametric effect of Context indicates that the median curve for targets after Agreeing determiners was lower than after Non-Agreeing determiners ($\beta = -0.814$, $p < .001$). In other words, participants looked more towards the target in Non-Agreeing sentences like (4), which used the determiner *the* (bottom right), than in Agreeing sentences like (3), where the target was preceded by an agreeing determiner (top right). Finally, the interaction between Match and Context indicates that the Match effect was smaller in Agreeing sentences than in Non-Agreeing sentences ($\beta = -0.565$, $p < .001$).

Table 5: Model summary of target advantage for looks to singular syll-targets. Parametric intercept represents the overall median. The 'baseline' smooth represents the curve with the Match value of Match (stem duration is lengthened to match expected unsuffixed duration), and the Context value of Agreeing (preceded by an agreeing determiner). Following smooths reflect differences between the baseline smooth and the other combinations of Match and Context.

Parametric terms				
Term	**Estimate**	**Std. Error**	***z***	***p***
(Intercept)	1.526	0.282	5.405	<.001
Match	1.812	0.073	25.877	<.001
Context	−0.814	0.070	−11.587	<.001
Match × Context	−0.565	0.139	−4.057	<.001
Smooth terms:				
Term	**Effective df**	**Ref.df**	χ^2	***p***
Baseline	4.98	5.56	49.35	<.001
Match = Mismatch, Context = Agreeing	4.98	6.02	89.16	<.001
Match = Match, Context = Non-Agreeing	4.05	4.96	25.15	<.001
Match = Mismatch, Context = Non-Agreeing	4.37	5.33	76.21	<.001
(by-subject smooths)	267.55	509.00	2070.96	<.001
(by-item smooths)	531.31	909.00	6854.40	<.001

the Mismatch curve for Non-Agreeing syll-targets – which is exactly the combination of conditions that produced the most symmetrical distribution of responses (**Figure 3**, bottom right panel). Plots of the model predictions with raw data overlaid can be seen in the Supplementary Materials.

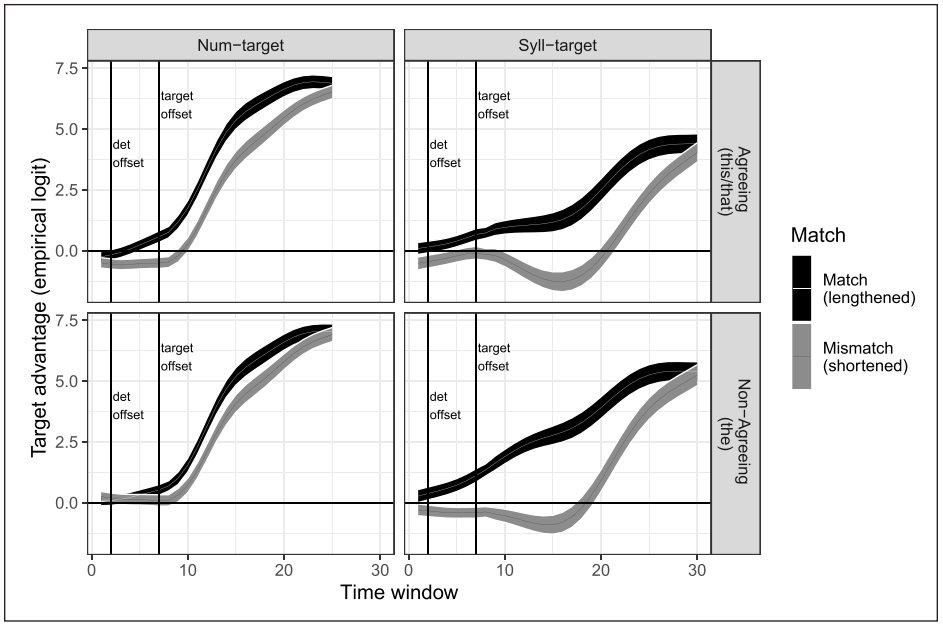

Figure 6: Model estimates of median gaze trajectories to num-targets (left) and syll-targets (right) across Agreeing (top) and Non-Agreeing (bottom) contexts. Lines show model estimates for each time window, while ribbons represent 95% confidence intervals of model estimates. Targets in the Match condition (solid white lines with black ribbons) showed higher median target advantage than targets in the Mismatch condition (dotted black line with gray ribbons).

The smooth terms indicate that the curve shapes for the default condition – Matching duration and Agreeing Context – were significantly different from the other three combinations of conditions. As **Figure 6** illustrates, the curve for the default condition (top right, black ribbon) rises slowly until time window 17 or so, before rising more steeply thereafter. In the Non-Agreeing condition, by contrast, the Match curve (bottom right, black ribbon) rises much more steadily from the target offset. The two Mismatch curves (right panels, gray ribbons) have quite decidedly distinct shapes compared to the Match-Agreeing reference curve. After the offset of the target, both curves show a distinct dip in target advantage, reflecting the period when participants would be misled into looking toward the competitor. At about time window 18, the Mismatch curves rise steeply, as participants recover and find the target. However, in the Non-Agreeing condition, the confusion from the Mismatch curve seems reduced relative to the Agreeing condition: The dip after target offset is less extreme, and the recovery happens sooner, crossing the 0-line into positive target advantage two time windows – roughly 100 ms – earlier.

3.2.2 Num-targets

The model summary for num-targets is shown in **Table 6**. The parametric effects showed a significant interaction between Match and Context, while the smooth terms suggest that

differences in curvature across the four conditions reflect only the Match manipulation. In the parametric terms, the main effect of Match ($\beta = 0.878$, $p < .001$), and the main effect of Context ($\beta = -0.205$, $p = .004$) went in the same direction with num-targets as with syll-targets. Participants showed higher overall target advantage when stem duration was lengthened in the Match condition (black ribbons, left side of **Figure 6**) than when it was shortened in the Mismatch condition (gray ribbons). They showed lower target advantage following the demonstratives *this/that* in the Agreeing condition (top left of **Figure 6**) than following *the* in the Non-Agreeing condition (bottom left). However, the interaction between Match and Context went in the opposite direction from syll-targets. While syll-targets showed a smaller Match effect in Agreeing sentences, num-targets, by contrast, had a larger Match effect in Agreeing sentences ($\beta = 0.392$, $p = .005$). This can be observed by comparing the top and bottom panels on the left side of **Figure 6**, which show a larger difference between black Match and gray Mismatch ribbons in the top panel of Agreeing sentences than in the bottom panel of Non-Agreeing sentences.

Table 6: Model summary of target advantage for looks to singular num-targets. Parametric intercept represents the overall median. The baseline smooth represents the curve with the Match value of Match (stem duration is lengthened to match expected unsuffixed duration), and the Context value of Agreeing (preceded by an agreeing determiner). Following smooths reflect differences between the baseline smooth and the other combinations of Match and Context.

Parametric terms				
Term	**Estimate**	**Std. Error**	***z***	***p***
(Intercept)	3.085	0.233	13.243	<.001
Match	0.878	0.069	12.599	<.001
Context	−0.205	0.070	−2.920	.004
Match × Context	0.392	0.139	2.827	.005
Smooth terms:				
Term	**Effective df**	**Ref.df**	χ^2	***p***
Baseline	6.61	7.27	196.99	<.001
Match = Mismatch, Context = Agreeing	3.29	4.06	24.84	<.001
Match = Match, Context = Non-Agreeing	2.21	2.75	3.16	.254
Match = Mismatch, Context = Non-Agreeing	3.50	4.31	37.00	<.001
(by-subject smooths)	266.353	509.00	2341.09	<.001
(by-item smooths)	456.271	909.00	2915.34	<.001

The smooth terms show that there was an overall difference in curvature between Match and Mismatch gaze traces, reflecting the earlier onset of the rise to peak target advantage for Match targets, compared to a delayed onset of the rise for Mismatch targets. However, the Match curve shapes are quite similar across Agreeing and Non-Agreeing contexts.

4. Discussion

This experiment was designed to test whether listeners are ideal observers, highly attentive to detailed phonetic information, or whether they are more strategic, employing a utility function to adjust how they attend to cues of differing usefulness. Baseline ideal observers should always attend to noun stem duration as a cue to whether a plural suffix follows, even if that information is rendered redundant by the presence of a preceding determiner that signals the number of the following noun. This would appear in the results as an additive, non-interactive effect of Match, such that Match trials would show more target advantage than Mismatch trials, regardless of sentence context. Phonetic predictors should show stronger benefits of Match after Agreeing determiners, because the determiners allow them to form detailed expectations about the form of the noun stem, allowing them to process the stem faster when it matches those expectations. This was predicted to emerge as an interaction between Match and Context, with stronger effects of Match after Agreeing determiners than after Non-Agreeing determiners for num-targets.

If, on the other hand, listeners are strategic, then they should down-weight rapidly changing acoustic information, because forming rapid real-time predictions with that information is too costly when it is made redundant by morphosyntactic context. This perceptual strategy would also result in an interaction between Match and Context with num-targets, but in the opposite direction from phonetic predictors. The target advantage should be smaller in Agreeing contexts, where the agreeing determiner renders stem duration redundant, and larger in Non-Agreeing contexts, where the absence of morphosyntactic information means that stem duration carries more information.

In the syll-targets, none of these accounts predicts any interaction with Context, because the presence or absence of an agreeing determiner is irrelevant to distinguishing *rock* from *rocket*, as both are singular nouns.

The results for accuracy and reaction time are not consistent enough to draw any strong conclusions. For accuracy, there was a tendency for Match targets to elicit more accurate responses than Mismatch targets, but only with syll-targets. This corroborates previous findings focused on polysyllabic shortening (Blazej & Cohen-Goldberg, 2015; Davis et al., 2002; Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005; Salverda et al., 2003; Shatzman & McQueen, 2006), but does nothing to extend them to segmental compression, or shed light on the three accounts tested here. For reaction time, Match targets elicited significantly faster responses than

Mismatch targets, an effect which applied to both num-targets and syll-targets, and was larger for the latter. This suggests that listeners do use segmental compression patterns to aid perception, albeit to a lesser extent than polysyllabic shortening. The lack of any interaction between Match and Context points toward the baseline ideal observer account.

Yet these patterns of results do not tell the whole story, for a variety of reasons. First, the accuracy of responses was quite high, especially for num-targets. This close to ceiling, it is difficult to extract much nuance from the distribution of a handful of incorrect responses. Second, neither accuracy nor latency of the responses directly reflects online processing of phonetic detail in real time. These measures instead reflect a decision-making process that incorporates not only phonetic detail, but also the following sentence context with its additional disambiguating information. Participants' responses are further filtered through the opportunity to consciously consider their interpretation of the stimulus, followed by the programming and execution of the motor response of the right hand. The fact that any Match effect at all emerged from such a crude measure of auditory perception – especially for num-targets, where sensitivity to segmental compression effects has not been demonstrated in perception yet – is itself a novel finding. However, to draw conclusions about the actual mechanism of online perception, it is necessary to consider the gaze traces.

4.1 Num-targets: Evidence for phonetic prediction

The gaze traces for num-targets showed a significant Match effect – the first demonstration that listeners can attend to segmental compression as well as polysyllabic shortening in processing morphologically complex words. The magnitude of the effect, however, was weaker in num-targets than in syll-targets, despite the identical durational manipulation. This could reflect the fact that the contrast between a singular and a plural, which both belong to the same lexeme, is less important for communication than the contrast between singular and two-syllable words, which belong to entirely separate lexemes. Alternatively, it may reflect the fact that contrasts between singulars and plurals can be reinforced by other morphosyntactic cues in a given sentence. By contrast, the difference between a two-syllable word and a one-syllable word lexically embedded within it may rest entirely on the phonetic information contained within the words themselves, leading listeners to attend more diligently. Or, perhaps, they simply benefit from the greater amount of time affforded by the extra syllable to process the phonetic detail more fully. Regardless, the presence of the Match effect in num-targets confirms that listeners' sensitivity to morphologically-conditioned phonetic detail operates at very fine scales, within the space of a single syllable.

This finding complements and expands upon previous work showing listeners' use of morphologically-conditioned durational variation – in particular, with affixes. Schmitz (2022),

for example, examined listeners' attentiveness to the duration of word-final /s/, which tends to be longer when it is part of a monomorphemic word, such as *mix*, than when it is a plural suffix, as in *books*. (See also Plag et al., 2017; Zimmermann, 2016, although cf. Seyfarth et al., 2018.) Using mouse-tracking in a number-decision task, Schmitz (2022) observed that listeners moved the cursor more directly across the screen toward the correct answer when the duration of word-final /-s/ matched these tendencies – longer for monomorphemic words and shorter for suffixed words – than when it reversed them.

Research on prefixes, too, has shown similar listener sensitivity to this sort of fine phonetic detail, in the context of words like *discolor* and *mistime*. According to R. Smith & Hawkins (2012), the meaning of these words is the composition of the meaning of the stem and the negating meaning of the prefix, and so these words can be characterized as containing a true prefix before the stem. Words like *discover* and *mistake* cannot be so decomposed; and so even though they contain the same initial triphones *dis-* and *mis-*, they are better described as pseudo-prefixed. The pronunciation of true prefixes and pseudo-prefixes varies systematically across a number of phonetic parameters, including duration. Vowels in true prefixes are longer than in pseudo-prefixes, while final consonants are shorter (R. Smith & Hawkins, 2012). Clayards et al. (2021) examined perception of these prefixes, along with two more – *re-* (as in *re-strings/restricts*), and *ex-* (as in *expletive/ex-policemen*) – and reported that, in addition to the other phonetic patterns reported by R. Smith & Hawkins (2012), true prefixes are longer than pseudo-prefixes overall, and that listeners can use these phonetic patterns to guide online speech perception of pseudo-prefixed and true-prefixed words.

The current study builds on these findings in two ways. First, the results reported in Clayards et al. (2021) are subject to an unavoidable constraint imposed by the vocabulary of English: There are only so many true/pseudo-prefix pairs in the language, and it is difficult to come up with many word pairs while respecting the strict controls the authors placed on the phonemic overlap and prosodic structure of the paired stimuli. By contrast, the current study used noun stems, which are far more numerous than syllables with a dual identity as true prefixes and pseudo-prefixes. As a result, the findings reported here provide evidence for listeners' use of these durational patterns across scores of words in real-time speech perception.

Second, both Schmitz (2022) and Clayards et al. (2021) created their matching and mismatching stimuli by cross-splicing affixes. This approach preserves all components of fine phonetic detail associated with the donor words, faithfully capturing the richness of pronunciation variation that listeners have at their disposal. Yet such an approach makes it difficult to determine how abstract listeners' knowledge of these pronunciation patterns might be. Do listeners need the full range of subphonemic cues to guide their perception of morphological structure, or can one cue in isolation do the trick? The current study did not cross-splice the stimuli, but instead manipulated duration alone, and so reveals that listeners do not require the full set of fine phonetic detail to

discriminate morphological structure. Although cross-splicing may well produce a larger Match effect, the current study shows that duration of the stem by itself allows listeners to better distinguish unsuffixed from suffixed words, suggesting a degree of abstract generalization in their use of phonetic cues during speech perception.

Crucially, the Match effect for num-targets interacted with Context, in a direction that exactly matches the phonetic predictor account. As visualized on the left side of **Figure 6**, num-targets enjoyed a larger Match effect (top) after Agreeing determiners than after Non-Agreeing determiners (bottom) This is incompatible with the strategic listener account, which predicts the opposite pattern. It is also incompatible with the baseline ideal observer account, which predicts no difference across contexts. Even the timing of the separation between Match and Mismatch follows the mechanism of the phonetic predictor account: the separation occurs earlier after Agreeing determiners than Non-Agreeing determiners, which is expected if listeners are able to make predictions earlier based on the preceding morphosyntactic context.

If listeners' ability to predict the number of the num-target is responsible for the strengthening of the Match effect, then it raises the intriguing possibility that being able to predict the identity of the syll-target might also increase the magnitude of the Match effect, by the same mechanism. Such a manipulation has not, to my knowledge, been investigated. In the existing studies that demonstrated robust responsiveness to polysyllabic shortening patterns in sentence contexts (Blazej & Cohen-Goldberg, 2015; Davis et al., 2002; Salverda et al., 2003; Shatzman & McQueen, 2006), the experimenters took great care to ensure that the sentences up to the target provided only neutral semantic context. Listeners could not use that preamble to predict whether the free-standing word (*cap*) or the two-syllable word (*captain*) was more likely. Suppose, though, that the sentences *did* provide adequate semantic context to allow listeners to predict the identity of the target word. If so, then part of that prediction would include the polysyllabic shortening pattern associated with a two-syllable word, or the lengthening pattern associated with the single-syllable stem. We might therefore expect that listeners would show larger effects of polysyllabic shortening in contexts where the syll-target's identity can be predicted, by the same mechanism in the current findings that produced larger effects of segmental compression in contexts where the num-target's number could be predicted.

4.2 Syll-targets: Difficult demonstratives?

The results for syll-targets (**Figure 6**, right-hand side) robustly replicated previous findings showing listener sensitivity to polysyllabic shortening (Blazej & Cohen-Goldberg, 2015; Davis et al., 2002; Kemps, Ernestus, et al., 2005; Kemps, Wurm, et al., 2005; Salverda et al., 2003; Shatzman & McQueen, 2006). Mismatch gaze traces hover around a target advantage of 0 until after the target offset. Then they dip, reflecting the temporary advantage of the competitor, which is, at that moment, the best match to the shortened stem in its intentionally ambiguous phonetic

context. In other words, when listeners heard *rock attracts*, the shortening of the stem, combined with the initial phones of *attract*, pulled their gaze more toward the competitor *rocket* than towards the target *rock*. In the Match condition, the lengthening of the stem shielded listeners from that initial misparse, and target advantage rises consistently. The rise is more gradual with syll-targets than with num-targets, but that is unsurprising, as syll-targets were embedded in sentences whose following contexts were intentionally ambiguous, while num-targets were not.

A complication in the syll-target results is that, although the Match effect interacts with Context as it did with num-targets, it interacts in the opposite direction. Match targets have a higher target advantage over Mismatch targets in the Non-Agreeing sentences than in the Agreeing sentences. Yet this cannot be interpreted as evidence against the phonetic predictor account, or indeed evidence for the strategic listener account, because with syll-targets, both target and competitor were singular. Any agreeing determiner will be singular, and hence be equally compatible with both target and competitor, in exactly the same way a non-agreeing determiner is. The difference in the Match effect, therefore, cannot stem from any difference in suitability between determiner and target or determiner and competitor. What is responsible for it?

One possibility has to do with the nature of the determiner itself. Recall that, in addition to the weakening of the Match effect, syll-targets in the Agreeing sentences also showed a lower overall target advantage than their counterparts in the Non-Agreeing sentences. This suggests that demonstrative determiners might be, overall, more difficult for listeners to process than the definite determiner. Indeed, one pilot participant reported feeling that sentences with demonstrative determiners felt subjectively harder to understand than those with only definite determiners.

What is it, then, that makes demonstratives more difficult? Perhaps it is the contextual licensing that governs the form they take. A definite determiner has only one form in English – *the*. By contrast, even leaving aside the number agreement, demonstrative determiners can take a proximal form (*this/these*) or a distal form (*that/those*). The proximal/distal contrast carries with it substantial discourse-pragmatic information, which can reflect a variety of factors, including the location of the referent involved; its discourse status, and whether that status is common ground between speaker and hearer, or privileged only to the speaker; and sometimes even physical properties of the referent, such as size and harmfulness (Peeters et al., 2021). However, the sentences in this study were presented in isolation, lacking any of the real-world or discourse-pragmatic context that could license the choice of demonstrative. As a result, in sentences where any morphosyntactic agreement information was valueless, listeners may have struggled to process the demonstratives more than the definite determiner *the*. If recovering from this struggle spilled over into the following noun, then that could account for the weaker target advantage for

syll-targets in the Agreeing sentences, where they followed demonstratives, relative to the Non-Agreeing sentences, where they followed *the*.

It is this recovery period that could also be responsible for weakening the Match effect. Recall that cognitive load reduces listeners' ability to process phonetic detail (Christiansen & Chater, 2016; Mattys & Wiget, 2011; Mattys et al., 2009, 2014). Although the work by Mattys and colleagues imposed a cognitive load by virtue of a secondary, non-linguistic task, it is possible that the demands imposed by processing unlicensed demonstrative determiners had a similar impact. With their cognitive capacity otherwise occupied, listeners could not spare the same resources to handle phonetic detail the way they could following straightforward, easy-to-process *the*. As a result, the Match effect was reduced.

This interpretation has some conceptual similarities to the strategic listener account, because it suggests that there are indeed contexts where limits on cognitive capacity reduce listeners' attention to phonetic detail. The key difference, however, is that it is not a strategic response to the informativity or redundancy of that detail. The Match manipulation is equally useful in distinguishing syll-target from competitor regardless of the nature of the preceding determiner. Rather, the reduced use of phonetic detail seems to be an automatic consequence of the increased processing load. To the extent that listeners are dynamically adjusting their use of phonetic detail, they may not be doing it out of strategy, but out of necessity.

4.3 Agreement vs. discourse-pragmatic licensing

If unlicensed demonstrative determiners are indeed more difficult to process than the definite determiner, then comparing their effect on syll-targets and num-targets affords an opportunity to evaluate the strength of agreement cues against discourse-pragmatic cues. In languages whose determiners regularly mark agreement, such as French, Spanish, and German, listeners show an advantage in identifying targets whose determiners disambiguate them from competitors (Berends et al., 2016; Dahan et al., 2000; Hopp, 2013; Lew-Williams & Fernald, 2007). Dahan et al. (2000), for example, asked listeners to distinguish a target like *le bouton* 'the.MASC button' from a cohort competitor that began with the same few segments, but differed in gender, as in *la bouteille* 'the.FEM bottle'. In French, definite determiners agree in gender before singular nouns, but before plural nouns take the same form – *les* – regardless of gender. Listeners were able to distinguish the target from the competitor more effectively when it was preceded by a singular definite determiner, which provided gender information, than when it was preceded by a plural definite determiner, which did not. Lew-Williams & Fernald (2007) showed a similar pattern in Spanish, which was present in children as young as 34 months. Listeners distinguished targets from competitors more quickly when they differed in gender than when they shared a gender,

because the agreement information on the determiner allowed listeners to exclude the competitor from consideration immediately. Hopp (2013) extended this pattern to German, demonstrating that not only can adults use gender-agreeing determiners to distinguish targets from competitors, but so can L2 speakers.

Even in English, which lacks grammatical gender, verb agreement in questions can be used to predict the identity of following subject nouns. Lukyanenko & Fisher (2016) showed adults and three-year-old children pictures of a singular or plural target, and asked them to respond to sentences whose verbs were either informative about target number (*Where is/are the good cookie(s)?*) or uninformative (*Can you find the good cookie(s)?*). They found that both adults and children began to look to the target picture even before the noun onset, as long as they had the informative agreement about target number in the verb. With uninformative instructions, they had to wait until the target noun was produced before they could look to the correct picture. Brown et al. (2022) replicated this finding in adults, and further showed that, in addition to directing listeners' attention to the target object, verb agreement information also reduced looks to similar-sounding competitors that did not match the target's number.

Clearly, then, morphosyntactic agreement is a robust and useful cue in online sentence processing. Is it strong enough to overcome the disadvantage posed by processing pragmatically complex demonstrative determiners? Although the design of this study was not constructed to directly test this question, it does accidentally offer some evidence that the answer is no. The main parametric effect of Context for num-targets was negative, indicating an overall disadvantage for num-targets in the Agreeing sentences (with demonstrative determiners) relative to the Non-Agreeing sentences (with *the*). In other words, whatever benefit was conferred by the agreement information in the demonstratives for distinguishing num-targets from plural competitors, it was not enough to overcome the difficulty associated with demonstratives relative to *the*. Nevertheless, the benefits of that agreement information were still detectable. The size of the Context effect – i.e., the disadvantage associated with demonstratives – was substantially smaller with num-targets, coming in at -0.205 elog units of target advantage, than with syll-targets, whose disadvantage, at -0.814 elog units, was nearly four times as large.

Figure 7 visualizes this result, collapsing across the Match and Mismatch conditions. The disadvantage brought by demonstrative determiners in the Agreeing context is clearly visible in the right-hand pane, which shows gaze traces for syll-targets. In the left-hand pane of num-targets, however, where the agreement information in the demonstrative determiners has the opportunity to offset the disadvantage, that difference is greatly reduced.

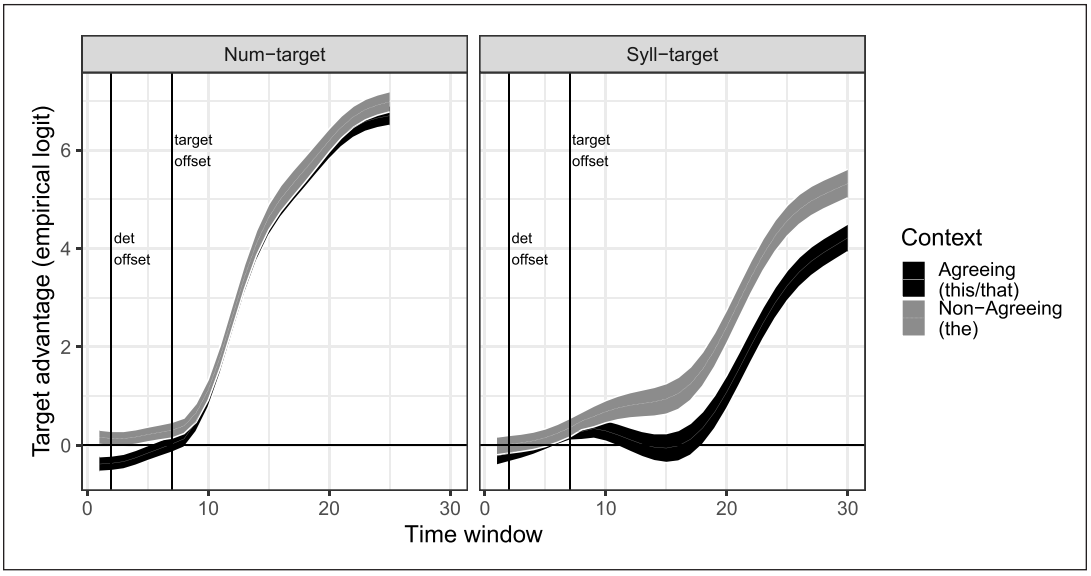

Figure 7: Modeled gaze-traces aggregated across Match and Mismatch, comparing the effect of Context for num-targets (left) and syll-targets (right).

5. Conclusion

Listeners are able to attend to fine phonetic detail, even in contexts where it is informationally redundant. They can use their knowledge of polysyllabic shortening patterns to resolve temporary ambiguities introduced in cases of lexical embeddings, and this study reveals that they can use their knowledge of segmental compression to get a head start on morphosyntactic processing. In particular, the current findings show that this use of segmental compression is a generalization of an abstract pattern: Listeners did not draw on lexically specific fine phonetic detail, but rather responded to a purely durational manipulation.

Furthermore, listeners' ability to use that information reflects the context and processing demands imposed by the rest of the sentence. When the context allows listeners to hone their predictions about the phonetic realization of upcoming information even a little, they will so hone them. However, when the context imposes a processing demand that is not offset by any added fine-tuning of predictions, listeners will reduce their attention to that phonetic detail. The role of extrinsic cognitive load in restricting attention to phonetic detail is well established in perception of isolated words (Mattys & Wiget, 2011; Mattys et al., 2009, 2014), but this study is the first to show that it may be imposed by the linguistic structure of the sentence context itself. Research is in progress to follow up on this possibility.

The findings of this work are broadly consistent with the predictions of ideal observer models, which hold that listeners observe and incorporate all information into their ongoing incremental

interpretation of a speech input as it unfolds in real time, but, crucially, only when they can spare the cognitive capacity to do it. When that capacity is taxed by processing difficult structures, listeners lose just a bit of their ability to attend to low-level phonetic cues. Their observational skills are close to ideal, but nothing is without limit.

Appendix: Stimuli

Table 7: Singular targets and two-syllable competitors, along with sentence frames. Plural competitors were simply the plural versions of the singular targets. Underlines indicate the phonemic contexts following target words that were constructed to mimic the second syllable of two-syllable counterparts.

Singular	Two-Syllable	Sentence
ant	antler	The rock pile protected this ant lurking in the undergrowth
arm	armour	The advertisement showed this arm around a striped pole
awl	olive	I won't give you this awl if you're not going to use it
bag	baggage	The designer made this bag adjustable and customisable
band	bandage	We want this band adjusted to match the size of the box
bar	barley	He asked if this bar legally belonged to him
beak	beaker	The professor damaged this beak early in the semester
bee	beaver	I did the project on this bee virtually on my own
bell	belly	Keep an eye on that bell even if you don't think it will ring
bill	building	The magnitude of this bill diminished your savings substantially
bowl	boulder	The combination of minerals makes that bowl durable and strong
bride	bridle	You'll need to treat this bride a little like a queen
buck	bucket	We saw that buck at the trail junction
bud	butter	We wondered whether to put this bud or the fern in the vase
bun	bunny	The appearance of that bun easily overcame my dieting willpower
cab	cabbage	If you need this cab adjust your expense account accordingly
can	candle	I found the story about that can duller than dishwater
cap	capsule	After being dunked in liquid nitrogen this cap simply shattered
car	cartridge	Using this car truly would make the deliveries go faster
card	cardinal	The beauty of that card nearly left me in tears
cart	carton	My view was blocked by that cart in front of the door
cheque	chequer	He offered me this cheque earnestly
core	coral	I found this core alone in the display case

(Contd.)

Singular	Two-Syllable	Sentence
cot	cotton	We quickly dug out this cot and a blanket when he arrived with a baby
crack	cracker	The sight of this crack arrested all traffic on the road
crow	crocus	We saw that crow collapse after flying into our glass door
cub	cupboard	The artist finished painting that cub around five in the evening
doll	dolphin	She made sure that that doll fit into her backpack
egg	exit	The museum showed this egg sitting next to the silver spoons
eye	iron	The children touched the painting of that eye right there in the middle
fan	phantom	I installed that fan to make my room less stuffy
fly	flyer	I first noticed this fly around three in the morning
foal	folder	I took a picture of that foal during our visit to the petting zoo
ham	hammer	The friend who gave me this ham arrived late to my dinner party
hand	handle	The metaphor of that hand elicits a sense of comradeship
harp	harpoon	He had already seen that harp oodles of times
hat	hatchet	The metal clip on that hat chipped my glasses
heel	helix	Fortunately this heel exactly matched the inseam of my best trousers
horn	hornet	I was not pleased to discover that horn attached to my seat with superglue
jack	jacket	The design of this jack attests to the age of the card deck
knee	needle	Flexibility in this knee delights any dance instructor
lamp	lamprey	The color of that lamp reminds me of my grandmother
lawn	laundry	The unexpected rain left this lawn dreary and unpleasant to sit on
lock	locker	He always uses this lock or a chain to secure his bike
mast	mastiff	The bad weather will be a problem for that mast if we don't repair it soon
monk	monkey	I didn't notice that monk even during our tour of the monastery
muff	muffin	I packed away this muff in mothballs for storage
net	nettle	If we're lucky this net 'll keep out badgers

(Contd.)

Singular	Two-Syllable	Sentence
pan	pansy	The scientist found on that pan zinc and aluminum from our local mine shafts
pane	painting	Behind this pane tin flower pots were visible
part	partridge	We need to make this part truly straight if the hairstyle is going to work
pen	pendant	She dropped this pen down the heating grate
pill	pillar	I need that pill early in the morning to ward off migraines
pin	pinto	I bought this pin to stock my emergency sewing kit
pit	pitcher	Unfortunately that pit choked him because he ate too fast
pole	poultry	We tripped over that pole trying to cross the farm yard
post	postage	I caught sight of that post adjacent to the bus stop
rack	racket	I finally found that rack at the back of the cupboard
raft	rafter	I took a picture of that raft around the time we went camping
ramp	rampart	The illustrator drew that ramp artistically
rib	ribbon	I damaged this rib in a car accident
road	rodent	According to our guidebook, that road enters the forest two miles ahead
robe	robot	The ad said that this robe offers the ultimate in comfort
rock	rocket	It seems this rock attracts migrating songbirds
rug	rugby	She asked us to sell this rug before the end of the week
seal	ceiling	I'm not so fond of that seal in comparison to the otters
seed	cedar	I discovered this seed around the back of the garden
shack	shackle	The foreboding appearance of that shack alarmed me
shell	shelter	She cleaned off that shell to reveal beautiful iridescent colors
skull	sculpture	The time I spent studying that skull prepared me for med school
sling	slinky	It seems like that sling keeps getting tangled
slip	slipper	I can't find that slip around any stores any more
sock	socket	The investigator discovered this sock attached to the trouser leg
splint	splinter	It was necessary for this splint to remain in place for two weeks
spring	sprinkler	The experiment with this spring clearly demonstrated Hooke's law
stick	sticker	The toddler waved that stick around to ward off mosquitos

(Contd.)

Singular	Two-Syllable	Sentence
stilt	stilton	I was surprised to discover this stilt and pogo stick in the garage
tie	tile	I tripped on that tie looking for my shoes last night
toe	toaster	I broke this toe sticking my foot under the bed
toy	toilet	The baby's fascination with that toy lets me see that he likes my gift
track	tractor	He warned us that this track turns muddy in the spring
trail	trailer	We took pictures of that trail around the picturesque lake
tube	tuba	He finally found this tube among the clutter in his junk drawer
wall	wallet	The colour of this wall attested to their love of pink

Data accessibility statement

All analysis code and data is available on the following OSF archive: https://osf.io/5zma7/?view_only=59664a282c8348058d9e004632bef4b2.

Ethics and consent

All experimental protocols were reviewed and approved by the University of Glasgow College of Arts Ethics Committee, application number 100210116. All participants gave informed consent before participating in the experiment.

Acknowledgements

Many thanks are due to my research assistants, Charilaos Votsis, Varshneyee Dutt, and Katharina Kolland. I am grateful to Matthew T. Carlson and Giuli Dussias for guidance and support during the development of previous stages of this work, and to Rachel Smith, Jane Stuart-Smith, and the Glasgow University Laboratory of Phonetics for suggestions, discussion, and feedback on this project. I am further grateful to Rachel Smith and Sarah Hawkins for useful feedback on this manuscript, and to Harald Baayen for invaluable guidance on the statistical analyses reported here. This work was supported by funding from the Strategic Resources Allocation Fund from the University of Glasgow's School of Critical Studies.

Competing interests

The author has no competing interests to declare.

Author contributions

Aside from minor details of stimulus preparation and parts of data collection delegated to research assistants, Clara Cohen was fully responsible for all elements of this work.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(38), 419–439. DOI: https://doi.org/10.1006/jmla.1997.2558

Baayen, R. H., Fasiolo, M., Wood, S., & Chuang, Y.-Y. (2022). A note on the modeling of the effects of experimental time in psycholinguistic experiments. *The Mental Lexicon*, *17*(2), 178–212. DOI: https://doi.org/10.1075/ml.21012.baa

Barden, K., & Hawkins, S. (2014). Perceptual learning of phonetic information that indicates morphological structure. *Phonetica*, *70*(4), 323–342. DOI: https://doi.org/10.1159/000357233

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. DOI: https://doi.org/10.1016/j.jml.2007.09.002

Bates, D., Alday, P. M., Kleinschmidt, D. F., Santiago Calderón, J. B., Zhan, L., Noack, A., Bouchet-Valat, M., Arslan, A., Kelman, T., Baldassari, A., Ehinger, B., Karrasch, D., Saba, E., Quinn, J., Hatherly, M., Piibeleht, M., Mogensen, P. K., Babayan, S., Holy, T., … Nazarathy, Y. (2024). MixedModels documentation. DOI: https://doi.org/10.5281/zenodo.596435

Berends, S. M., Brouwer, S. M., & Sprenger, S. A. (2016). Eye-tracking and the visual world paradigm. In M. S. Schmid, S. M. Berends, C. Bergmann, S. M. Brouwer, N. Meulman, B. J. Seton, S. A. Sprenger, & L. A. Stowe (Eds.), *Designing research on bilingual development: Behavioral and neurolinguistic experiments* (pp. 55–80). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-11529-0_5

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. DOI: https://doi.org/10.1137/141000671

Blazej, L. J., & Cohen-Goldberg, A. M. (2015). Can we hear morphological complexity before words are complex? *Journal of Experimental Psychology: Human Perception and Performance*, *41*(1), 50–68. DOI: https://doi.org/10.1037/a0038509

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer. DOI: https://doi.org/10.1097/aud.0b013e31821473f7

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). "How to milk a coat:" the effects of semantic and acoustic information on phoneme categorization. *The Journal of the Acoustical Society of America*, *103*(5 Pt 1), 2670–2676. DOI: https://doi.org/10.1121/1.422787

Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language*, *125*, 104334. DOI: https://doi.org/10.1016/j.jml.2022.104334

Brown, V. A., Fox, N. P., & Strand, J. F. (2022). "Where are the … fixations?": Grammatical number cues guide anticipatory fixations to upcoming referents and reduce lexical competition.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *48*(5), 643–657. DOI: https://doi.org/10.1037/xlm0001019

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, 1–52. DOI: https://doi.org/10.1017/S0140525X1500031X

Clayards, M., Gaskell, M. G., & Hawkins, S. (2021). Phonetic detail is used to predict a word's morphological composition. *Journal of Phonetics*, *87*. DOI: https://doi.org/10.1016/j.wocn.2021.101055

Cohen, C., & Carlson, M. (2024). Shifting between storage and computation in lexical retrieval: Evidence from pronunciation variation. In M. Schlechtweg (Ed.), *Interfaces of phonetics* (pp. 155–204). De Gruyter. DOI: https://doi.org/10.1515/9783110783452-006

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465–480. DOI: https://doi.org/10.1006/jmla.1999.2688

Davis, M. H., Marslen-Wilson, W. D., & 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*(1), 218–244. DOI: https://doi.org/10.1037/0096-1523.28.1.218

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670. DOI: https://doi.org/10.1177/0956797610384743

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., & Goude, Y. (2021). Qgam: Bayesian nonparametric quantile regression modeling in *R*. *Journal of Statistical Software*, *100*(9). DOI: https://doi.org/10.18637/jss.v100.i09

Fox, N. P., & Blumstein, S. E. (2016). Top-down effects of syntactic sentential context on phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 730–741. DOI: https://doi.org/10.1037/a0039965

Gahl, S., & Baayen, R. H. (in press). Time and thyme again: Connecting spoken word duration in English to models of the mental lexicon. *Language*.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology. Human perception and performance*, *6*(1), 110–125. DOI: https://doi.org/10.1037/0096-1523.6.1.110

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. DOI: https://doi.org/10.1037/0033-295X.105.2.251

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3–4), 373–405. DOI: https://doi.org/10.1016/j.wocn.2003.09.006

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*(1), 33–56. DOI: https://doi.org/10.1177/0267658312461803

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variation in speech processing* (pp. 145–165). Academic Press.

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology* (pp. 25–40). Oxford University Press.

Katz, J. (2012). Compression effects in English. *Journal of Phonetics*, *40*(3), 390–402. DOI: https://doi.org/10.1016/j.wocn.2012.02.004

Kemps, R. J. J. K., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005). Prosodic cues for morphological complexity: The case of Dutch plural nouns. *Memory & Cognition*, *33*(3), 430–446. DOI: https://doi.org/10.3758/BF03193061

Kemps, R. J. J. K., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, *20*(1/2), 43–73. DOI: https://doi.org/10.3758/BF03193061

Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. *Journal of Phonetics*, *81*, 100984. DOI: https://doi.org/10.1016/j.wocn.2020.100984

Klatt, D. H. (1975). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*(5), 1208–1220. DOI: https://doi.org/10.1121/1.380986

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological review*, *122*(2), 148–203. DOI: https://doi.org/10.1037/a0038695

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, *31*(1), 32–59. DOI: https://doi.org/10.1080/23273798.2015.1102299

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America*, *51*(6.2), 2018–2024. DOI: https://doi.org/10.1121/1.1913062

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 192–198. DOI: https://doi.org/10.1111/j.1467-9280.2007.01871.x

Lukyanenko, C., & Fisher, C. (2016). Where are the cookies? Two- and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition*, *146*, 349–370. DOI: https://doi.org/10.1016/j.cognition.2015.10.012

Mattys, S. L., Barden, K., & Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin & Review*, *21*(3), 748–54. DOI: https://doi.org/10.3758/s13423-013-0544-7

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*(3), 203–243. DOI: https://doi.org/10.1016/j.cogpsych.2009.04.001

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*(2), 145–160. DOI: https://doi.org/10.1016/j.jml.2011.04.004

Mirman, D. (2014). *Growth curve analysis and visualization using R*. CRC Press.

Mitterer, H., & Reinisch, E. (2017). Surface forms trump underlying representations in functional generalisations in speech perception: The case of German devoiced stops. *Language, Cognition and Neuroscience*, *32*(9), 1133–1147. DOI: https://doi.org/10.1080/23273798.2017.1286361

Munhall, K., Fowler, C., Hawkins, S., & Saltzman, E. (1992). "Compensatory shortening" in monosyllables of spoken English. *Journal of Phonetics*, *20*(2), 225–239. DOI: https://doi.org/10.1016/s0095-4470(19)30624-2

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. DOI: https://doi.org/10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. DOI: https://doi.org/10.1016/S0010-0285(03)00006-9

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. DOI: https://doi.org/10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. DOI: https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Peeters, D., Krahmer, E., & Maes, A. (2021). A conceptual framework for the study of demonstrative reference. *Psychonomic Bulletin & Review*, *28*(2), 409–433. DOI: https://doi.org/10.3758/s13423-020-01822-8

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7*(pp. 101–140). Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110197105.1.101

Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, & Psychophysics*, *78*(1), 334–345. DOI: https://doi.org/10.3758/s13414-015-0981-7

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, *53*(1), 181–216. DOI: https://doi.org/10.1017/S0022226715000183

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in american english. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303. DOI: https://doi.org/10.1121/1.1912974

Rohde, H., & Ettlinger, M. (2012). Integration of pragmatic and phonetic cues in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 967–983. DOI: https://doi.org/10.1037/a0026786

Saito, M., Tomaschek, F., Sun, C.-C., & Baayen, R. H. (2024). Articulatory effects of frequency modulated by inflectional meanings. In M. Schlechtweg (Ed.), *Interfaces of Phonetics*. De Gruyter.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51–89. DOI: https://doi.org/10.1016/S0010-0277(03)00139-2

Schmitz, D. (2022). *Production, perception, and comprehension of subphonemic detail: Word-final /s/ in English*. Language Science Press. DOI: https://doi.org/10.5281/zenodo.7267830

Seyfarth, S., Garellek, M., Gillingham, G., Ackerman, F., & Malouf, R. (2018). Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience*, *33*(1), 32–49. DOI: https://doi.org/10.1080/23273798.2017.1359634

Shatzman, K. B., & McQueen, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science*, *17*(5), 372–377. DOI: https://doi.org/10.1111/j.1467-9280.2006.01714.x

Siddins, J., Harrington, J., Kleber, F., & Reubold, U. (2013). The influence of accentuation and polysyllabicity on compensatory shortening in German. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August), 1002–1006.

Smith, L. B., & Colunga, E. (2012). Developing categories and concepts. In *Cambridge handbook of psycholinguistics* (pp. 283–310).

Smith, R. (2015, September). Perception of speaker-specific phonetic detail. In S. Fuchs, D. Pape, C. Petrone, & P. Perrier (Eds.), *Individual differences in speech production and perception* (pp. 11–38, Vol. 3). Peter Lang D. DOI: https://doi.org/10.3726/978-3-653-05777-5

Smith, R., & Hawkins, S. (2012). Production and perception of speaker-specific phonetic detail at word boundaries. *Journal of Phonetics*, *40*(2), 213–233. DOI: https://doi.org/10.1016/j.wocn.2011.11.003

Smith, R., & Rathcke, T. (2017). Glasgow gloom or Leeds glue? Dialect-specific vowel duration constrains lexical segmentation and access. *Phonetica*, *74*(1), 1–24. DOI: https://doi.org/10.1159/000444857

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction.

Suomi, K. (2007). On the tonal and temporal domains of accent in Finnish. *Journal of Phonetics*, *35*(1), 40–55. DOI: https://doi.org/10.1016/j.wocn.2005.12.001

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, *27*(7–8), 979–1001. DOI: https://doi.org/10.1080/01690965.2011.597153

Trude, A. M., Tremblay, A., & Brown-Schmidt, S. (2013). Limitations on adaptation to foreign accents. *Journal of Memory and Language*, *69*(3), 349–367. DOI: https://doi.org/10.1016/j.jml.2013.05.002

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (Eds.), *Methods in empirical prosody research* (pp. 1–28, Vol. 3). Walter de Gruyter. DOI: https://doi.org/10.1515/9783110914641.1

van Alphen, P., & McQueen, J. M. (2001). The time-limited influence of sentential context on function word identification. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1057–1071. DOI: https://doi.org/10.1037/0096-1523.27.5.1057

White, L., & Turk, A. E. (2010). English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics*, *38*(3), 459–471. DOI: https://doi.org/10.1016/j.wocn.2010.05.002

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36. DOI: https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized Additive Models: An introduction with R*. Chapman and Hall/CRC.

Zimmermann, J. (2016). Morphological status and acoustic realization: Findings from New Zealand English. In C. Carignan & M. D. Tyler (Eds.), *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology (SST-2016)* (pp. 201–204). Australasian Speech Science and Technology Association.