Contextual speech rate influences morphosyntactic prediction and integration

Greta Kaufeld[a,b]*, Wibke Naumann[a], Antje S. Meyer[a,c], Hans Rutger Bosker[a,c], & Andrea E. Martin[a,c]

[a]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; [b]International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands; [c]Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands

Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500AH Nijmegen, The Netherlands; corresponding author email: greta.kaufeld@mpi.nl

Contextual speech rate influences morphosyntactic prediction and integration

Understanding spoken language requires the integration and weighting of multiple cues, and may call on cue integration mechanisms that have been studied in other areas of perception. In an eye-tracking experiment (visual-world paradigm), we investigated the combination and integration of contextual speech rate (a lower-level, perceptual cue) and morphosyntactic knowledge (a higher-level, linguistic cue). We observed that participants used contextual rate information immediately, and interpret this as evidence of perceptual inference and the generation of predictions about upcoming morphosyntactic information. Moreover, we found that early rate effects remained active in the presence of later conflicting lexical information. This result demonstrates that (1) contextual speech rate functions as a cue to morphosyntactic inferences, even in the presence of subsequent disambiguating information; and (2) listeners iteratively use multiple sources of information to draw inferences and generate predictions during speech comprehension. We discuss the implication of these demonstrations for theories of language processing.

Keywords: language comprehension; speech perception; cue integration; morphology; agreement; speech rate normalisation

1. Introduction

Speech is an important part of human behaviour. From energy fluctuations in the air, we are able to infer complex meaning, acquire novel information, and experience rich emotions. Doing so requires us to minimally map the properties of the acoustic signal onto more abstract units, such as phonemes, morphemes, syllables, words, and sentences. Establishing this mapping between perception and meaning is, however, rarely straightforward, because the acoustic speech signal does not carry unambiguous, physically quantifiable markers for abstract, hierarchical linguistic units and structures. On top of that, it can contain multiple sources of noise, variation and uncertainty. How does the brain accomplish this ill-posed task of mapping the acoustic signal onto linguistic units and structures? One branch of speech perception models aiming to help answer this question is tightly linked to psychophysiological models of *cue integration.* The goal of the current study is to contribute to our understanding of language comprehension by examining how signal-based, perceptual (relative duration) cues and knowledge-based, linguistic cues (morphosyntactic cues to gender) are iteratively combined within such a framework of cue integration.

1.1 Cue integration as a mechanistic model for perception

Cue integration as a psychophysiological mechanism has been researched in depth in the fields of vision and multisensory perception. The underlying idea is that our perceptual experience of the world emerges from drawing inferences based on the synthesis of multiple incoming pieces of sensory information, or *cues* (Ernst & Bülthoff, 2004; Fetsch, DeAngelis, & Angelaki, 2013). A cue can, in principle, be "any signal or piece of information bearing on the state of some property of the environment" (Fetsch et al., 2013, p. 12) or "any sensory information that gives rise to a sensory estimate" (Ernst & Bülthoff, 2004, p. 163; see also their brief discussion of why defining a cue is so hard). Multiple cues to a specific percept are *combined* by means of summation and, to alleviate the sampling uncertainty arising from the fact that different cues may not be equally reliable in any given situation, *integrated* (or weighted) by means of normalisation. A cue's reliability in a given situation is thus encoded as its weight during the integration process. This can be formalised both as a linear operation (Equation 1), or in terms of Bayesian inference (see, for example, Fetsch et al., 2013, or Landy, Banks, and Knill, 2011, for a more detailed overview of the underlying computations). One of the most attractive aspects of cue integration as a model of perception is the neurological plausibility of its underlying computations: summation and normalisation have been proposed as canonical neural computations that the brain uses to solve problems across different brain regions, modalities and contexts (Carandini & Heeger, 1994, 2012).

[Insert Equation 1]

1.2 Speech perception as cue integration

Models related to cue integration have been proposed for phoneme categorisation as early as the 1970s (e.g., Oden & Massaro, 1978; Sawusch & Pisoni, 1974). More recently, C-CuRE (“Computing Cues Relative to Expectations”; McMurray & Jongman, 2011), a model of speech perception that takes context into account, has been proposed and investigated extensively (e.g., Apfelbaum, Bullock-Rest, Rhone, Jongman, & McMurray, 2014; McMurray, Cole, & Munson, 2011; Toscano & McMurray, 2015). In C-CuRE, acoustic cues are encoded relative to specific values that the listener expects in a given situation. Crucially, these expectations can be established and adjusted based on other acoustic cues. The basic computation behind C-CuRE is linear regression: Initial regression equations predicting specific cue values are established based on previous knowledge and contextual information. These regression functions are a formalisation of what McMurray and Jongman (2011) term “expectations”. Newly perceived cues are interpreted relative to these expectations by computing the variance of the perceived cue from the value predicted by the regression function. Note that this notion of “computing cues relative to expectations” bears striking similarities to the concept of computing prediction errors within a predictive coding framework (Toscano & McMurray, 2015).

Models such as C-CuRE propose different types of acoustic cues that are involved in making categorisation decisions on a phonemic level, and they make some predictions about how these cues interact amongst each other (e.g., McMurray &

Jongman, 2011; Toscano & McMurray, 2015). However, they do not go beyond acoustic cues, and they do not make predictions about how phoneme categorisation might tie into a framework of speech comprehension that takes higher-level language comprehension as the goal of the perceptual system. There is wide-spread evidence that phoneme perception can be influenced by higher-level non-acoustic cues (e.g., Connine & Clifton, 1987; Fox, 1984; Ganong, 1980; Martin, Monahan, & Samuel, 2017; Pitt & Samuel, 1993; Rohde & Ettlinger, 2012; van Alphen & McQueen, 2001), so any comprehensive model of speech comprehension has to account for the ways in which sensory, signal-based cues interact with morphosyntactic, lexical, pragmatic, and other knowledge-based information online (cf. Kaufeld et al., in press).

Notably, a cue-based model of word segmentation was proposed by Mattys, Melhorn, and White (2005): Based on a series of word detection experiments, they suggested a hierarchy of cues for word segmentation, where both signal-based and knowledge-based cues are taken into account by the language comprehension system. The model is organised into three tiers (Tier I: lexical tier; Tier II: segmental tier; Tier III: metrical prosodic tier), with cues from higher levels of the tier hierarchy (corresponding to lexical and contextual information) taking precedence over cues from lower levels (Mattys et al., 2005). Based on a set of experiments (Mattys, Melhorn, & White, 2007), the authors later updated their model to include a more dynamic, "graded" relationship between cues from different tiers. Especially this later model is very similar in idea to models of cue integration, where cues can be dynamically combined across levels of perceptual hierarchy. However, the model suggested by Mattys and colleagues (2005; 2007) focusses on word segmentation, leaving open the important question of how the comprehension system achieves *understanding* above and beyond segmenting the acoustic signal into words.

1.3 Language comprehension as cue integration

In an effort to synthesise principles from perception, speech processing, and neurophysiology, Martin (2016) proposed cue integration as a general mechanism for language processing on all levels, outlining how such a model can begin to explain all stages of language comprehension and production, from sensory processing to dialogue. In this model, functional equivalents to formal linguistic representations and higher-level meaning are inferred from sensory information by iteratively extracting, combining, and integrating relevant linguistic cues (cf. Figure 1). Martin (2016) suggests a cascading architecture where cues can be combined and integrated across different levels of language comprehension through a process called sensory resampling. By resampling the input across different levels of processing, multiple cues can be derived from the same sensory input. Linguistic representations that have been inferred from sensory cues can thus, in turn, be cues for higher levels of representations. For example, acoustic cues can give rise to abstract percepts such as phonemes and morphemes; phonemes and morphemes can, in turn, act as cues towards the percept of a word; words can be cues to phrasal representations; and so on. In other words, cues are not only representations of the linguistic input, they also form the link between representations from different levels of linguistic hierarchy (Martin, 2016). Note how this differs from the notion of cues in most connectionist frameworks, such as the Competition Model (e.g., Bates & MacWhinney, 1987), where cues and their weights arise from inherent properties and features of a language. Within the model of cue integration suggested by Martin (2016), cues mark the transform of the sensory signals of speech and sign into structured linguistic representations. A significant part of this neural transform is performed by *internally generated* representations that have been generalized into linguistic knowledge after learning – potentially, but not exclusively, from language-inherent features.

[Figure 1]

Figure 1. Simplified graphical representation of the cue integration architecture for speech comprehension (adapted from Martin, 2016). Cues and their corresponding reliabilities, represented by Gaussian icons, are integrated across different levels of linguistic hierarchy. Predictions about upcoming linguistic information are visualized by black arrows pointing forward. Grey arrows represent sensory resampling, such that cues from different linguistic levels of representations can influence each other.

More generally, within a framework of cue integration, a psycholinguistic cue can be any source of information that is relevant for language processing, including endogenously generated representations and predictions (Martin, 2016). In the following section, we will briefly discuss how cue integration as a model of language comprehension can speak to the current debate about the role of prediction and anticipatory language processing (cf. Huettig, 2015; Nieuwland et al., 2018).

1.4 Cue integration and prediction during language processing

The role of our expectations about upcoming linguistic information in language comprehension has been investigated extensively in the last two decades (see Huettig, 2015; Nieuwland et al., 2018, for comprehensive reviews). Anticipatory language processing has been shown to occur for features on multiple levels of language processing, including semantic (Altmann & Kamide, 1999; Federmeier & Kutas, 1999; Federmeier, McLennan, Ochoa, & Kutas, 2002; Szewczyk & Schriefers, 2013), orthographic (Laszlo & Federmeier, 2009), morphosyntactic (Kamide, Scheepers, & Altmann, 2003; Kaufeld et al., in press; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003; Wicha, Moreno, & Kutas, 2003, 2004), and specific visual features (Rommers, Meyer, Praamstra, & Huettig, 2013). Based on these findings, several psycholinguistic models have been built on the assumption that prediction is one of the fundamental mechanisms of language processing (e.g., Dell & Chang, 2013; Pickering & Garrod, 2007). These models are in line with more general models of cognition where brains are seen as “prediction machines" that are "constantly engaged" in the task of minimizing the prediction error between incoming sensory information and previously established expectations (Clark, 2013). However, as Huettig and Mani (2016) and others (e.g., Huettig, 2015; Nieuwland et al., 2018; Rabagliati & Bemis, 2013) have pointed out, these "strict prediction models" fail to explain how we understand language in situations where upcoming linguistic information cannot (or need not) be predicted. In order to account for all of the available empirical findings, psycholinguistic models are needed that allow listeners to *make predictions when they can* (because it might be helpful for further language processing), but to *avoid doing so when they can't* (because the input might be too noisy or not informative enough).

As Martin (2016) points out, this optional capacity to make predictions can be implemented within a framework of cue integration. Bottom-up activity corresponds to integrated cues and their reliabilities, which are compared against top-down predictive activations. The potential mismatch between integrated cues and predictions is fed forward as a subset of cue reliabilities, corresponding to the notion of a "prediction error". Note that this ties in directly with the iterative nature of cue integration: The predictive activation itself acts as a cue for further processing and is therefore associated with a specific cue reliability (and thus weight) of its own, which is normalised against the reliability of all other available and relevant cues. Crucially, predictive activation does not necessarily *have* to occur: If the available lower-level cues to base predictions on are not reliable enough, or simply too sparse, no anticipatory language processing will be initiated.

1.5 Current study

In the current experiment, we asked how the speech comprehension system takes up and integrates cues from different levels of linguistic hierarchies, aiming to test predictions of the cue integration model as suggested by Martin (2016). More specifically, we asked three questions: First, does the system immediately use lower-level perceptual cues online in order to infer higher-level cues, even in the presence of subsequent disambiguating information? Second, are inferential gender cues immediately deployed to make predictions about upcoming linguistic information? Third, how does the system handle incoherence between inferences made based on an early perceptual cue and subsequent lexical information?

To address these questions, we conducted an eye-tracking experiment using the visual world paradigm. In the following two sections, we will briefly discuss two cues which will form the basis of our experiment. Contextual speech rate is a perceptual cue that has been argued to influence the earliest stages of phoneme categorisation; gender morphology is a linguistic cue that has been shown to influence linguistic prediction and integration. These two cues occur on different levels of linguistic hierarchy, so they will allow us to investigate cross-level integration online.

1.5.1 Contextual speech rate: An early perceptual cue

Contextual rate manipulations have been shown to influence duration-based phoneme perception: For instance, the perception of a vowel that is ambiguous between short /ɑ/ and long /a:/ in Dutch is biased towards /a:/ when embedded in a fast context sentence, but biased towards /ɑ/ when presented after a slow context sentence (e.g., Bosker, 2017a; Bosker, 2017b; Bosker & Reinisch, 2017; Bosker, Reinisch & Sjerps, 2017; Kaufeld et al., in press; Maslowski, Meyer, & Bosker, 2018; 2019a). Similar findings have been reported for other (duration-cued) segmental distinctions, such as /b-p/ (Gordon, 1988), /b-w/ (Miller & Baer, 1983; Wade & Holt, 2005), /p-p#p/ (Pickett & Decker, 1960), and singleton-geminate (Mitterer, 2018). In fact, reduced highly coarticulated linguistic units can even be missed entirely by listeners when presented in slow contexts. For instance, a reduced “terror” can be perceived as “tear”, omitting the second unstressed syllable “-or”, when embedded in a slow sentence (Baese-Berk et al., 2019). Similarly, the function word “or” in a phrase such as “leisure (or) time” can be perceived as present or absent depending on contextual speech rate (Dilley & Pitt, 2010), and the determiner “a” in a sentence such as *“The Petersons are looking to buy (a) brown hen(s) soon”* can perceptually “appear” or “disappear” when embedded in fast or slow contexts (Brown, Dilley, & Tanenhaus, 2012).

These effects of contextual speech rate are referred to by different names, such as “rate normalisation” (adopted here), “disappearing word effect”, “distal rate effect”, and “lexical rate effect” – but always involve rate-dependent speech perception. Rate normalisation effects have been observed to arise very early during perception, and they appear to modulate the uptake and weighting of other acoustic cues. Reinisch and Sjerps (2013) investigated the time course of the uptake and interplay of spectral, durational, and contextual cues for rate normalisation. Native speakers of Dutch were asked to categorise minimal word pairs such as /tɑk/ (*branch*) and /ta:k/ (*task*), where the vowel had been manipulated to be both spectrally and durationally ambiguous between /ɑ/ and /a:/, embedded in fast and slow contexts. They found that contextual rate cues were used very rapidly, influencing perception and categorisation of the target word at the same point in time as vowel-internal durational cues. These findings are in line with accounts of speech rate effects arising at early stages of lexical processing, potentially involving general auditory mechanisms (see also Bosker, 2017a; Bosker & Ghitza, 2018; Kaufeld et al., in press; Maslowski et al., 2019b; Wade & Holt, 2005; Sawusch & Newman, 2000; Miller & Dexter, 1988; but see Pitt, Szostak, & Dilley, 2016). Toscano and McMurray (2015) investigated the interplay of contextual rate effects with voice onset time (VOT) in an eye tracking experiment. English-speaking participants were asked to categorise minimal word pairs such as *beach* and *peach*, where the VOT of the initial plosive had been manipulated to be temporally ambiguous between the voiced and voiceless tokens. Eye gaze data indicated that contextual rate cues were used simultaneously with VOT cues, again suggesting that rate effects occur early during perception, and that contextual speech rate can be seen as a cue that modulates other acoustic cues. More recently, Kaufeld et al. (in press) assessed the flexible weighting and integration of morphosyntactic gender marking (a knowledge-based cue) and contextual speech rate (a signal-based cue). They reported robust speech rate normalization effects in participants' gaze patterns arising very early after vowel offset, even in the presence of preceding morphosyntactic information.

1.5.2 Gender morphology: A linguistic cue

There is plenty of evidence showing that listeners rapidly make use of morphological information during speech comprehension. In a corpus analysis of German, Bölte and Connine (2004) showed that gender-marked determiners can facilitate subsequent language processing, and gender priming effects have been reported for a multitude of languages, including German (e.g., Hillert & Bates, 1994; see Friederici & Jacobsen, 1999, for a comprehensive review of the gender priming literature). Importantly, gender information has been shown to influence both the prediction of upcoming referents (Szewczyk & Schriefers, 2013; Van Berkum et al., 2005; Wicha et al., 2004; but see Guerra, Nicenboim, & Helo, 2018; Kochari & Flecken, 2019), and the perception of following ambiguous phonemes (Martin et al., 2017).

1.5.3 Current experiment

In the current experiment, we examined the influence of contextual speech rate on the perception of the presence or absence of the morphosyntactic inflectional suffix /-ə/

(schwa), marking gender on indefinite determiners (feminine *eine* vs. neuter *ein*) in German. Participants were presented with two pictures on a screen, corresponding to a neuter and a feminine target noun (e.g., *Katze*$_{\textit{FEMININE}}$, “cat” vs. *Reh*$_{\textit{NEUTER}}$, “deer”), while listening to auditory instructions at fast or slow rates, asking them to look at one of the two pictures (*Schauen Sie jetzt sofort auf eine*$_{\textit{FEMININE}}$/*ein*$_{\textit{NEUTER}}$ *außergewöhnlich liebe*$_{\textit{FEMININE}}$ *Katze*$_{\textit{FEMININE}}$/*liebes*$_{\textit{NEUTER}}$ *Reh*$_{\textit{NEUTER}}$, “Now look at once at an$_{\text{FEMININE/NEUTER}}$ exceptionally friendly$_{\text{FEMININE/NEUTER}}$ cat$_{\text{FEMININE}}$/deer$_{\text{NEUTER}}$”). We had manipulated the indefinite determiner, *ein?*, to be ambiguous between perceived presence and absence (perceived either as *ein*, marking neuter, or as *eine*, marking feminine gender). Crucially, the indefinite determiner containing this ambiguous schwa phoneme was the earliest morphosyntactic cue indicating the gender (and, by proxy, lexical identity) of the target, thus allowing participants to make predictions about upcoming referents.

The cue integration model predicts that the system rapidly uses perceptual cues to draw inferences that are, in turn, deployed as cues for higher levels of processing. Previous findings reported by Brown et al. (2012) suggest that speech rate is, indeed, used by listeners to draw inferences about higher-level linguistic features, such as number. Brown and colleagues used a visual world paradigm to investigate listeners’ perception of the singular indefinite determiner “a” in a sentence such as *“The Petersons are looking to buy (a) brown hen(s) soon”*, where the carrier sentence surrounding the determiner region was manipulated to be either slow or fast. Overall, listeners were more likely to perceive the determiner as being “present” in fast as opposed to slow contexts, as evidenced by preferential looks towards pictures corresponding to a singular (plural) interpretation in fast (slow) contexts during the target time window. From a cue-integration perspective, this suggests that listeners used the acoustic cues from contextual speech rate and vowel duration to infer higher-level linguistic information about the number of the target noun.

In line with the findings reported by Brown et al. (2012) and the predictions from cue integration theory, we thus hypothesised that listeners would rapidly use lower-level contextual speech rate cues in order to infer higher-level morphosyntactic gender and lexical information. Specifically, when embedded in a fast context sentence, the ambiguous schwa phoneme should appear relatively long in contrast to the preceding phonemes – similar to more long /a:/ vowel responses after fast speech in Reinisch and Sjerps (2013). Participants should therefore be more likely to perceive the schwa as being present in a fast context, leading them to interpret the determiner as *eine*. This would, in turn, allow them to infer feminine gender based on the presence of the gender morpheme, and make predictions about the lexical identity of the target picture. Conversely, the ambiguous schwa phoneme should sound relatively short when embedded in a slow context sentence, possibly making the schwa perceptually disappear. Participants should thus be more likely to perceive the indefinite determiner as being *ein* in a slow context, allowing them to infer neuter gender and make predictions about the target picture's gender and lexical identity. Crucially, if participants used contextual rate cues to infer morphosyntactic information and then operationalised this information to make predictions about the target noun, we should find anticipatory looks to the relevant picture well before the onset of the target noun. Analysing a time window immediately after the onset of the ambiguous schwa phoneme and before target onset thus allowed us to address both the temporal (question 1) and the predictive aspect (question 2) of cue integration.

The cue integration framework predicts that cues can interact *across* levels of linguistic hierarchy – that is, signal-based, acoustic cues can influence the expectation and perception of knowledge-based, inferential cues, and vice versa (see also Mattys et al., 2007). However, to our knowledge, previous eye-tracking studies that investigated speech rate as a possible signal-based cue towards morphosyntactically relevant information have exclusively investigated it in combination with other ambiguous acoustic or morphosyntactic cues. That is, in Brown et al. (2012), the sibilant ("hen[*s* *s*]oon" vs. "hen [*s*]oon") was, itself, ambiguous in the experiment reported by Brown and colleagues. In fact, the authors specifically designed their stimuli to "increase participants' reliance on the determiner [...] as a cue to number" (Brown et al., 2012, p. 1375), and their analyses confirm that listeners based their judgements on the perception of the determiner, rather than a combination of both number cues (Brown et al., 2012). As such, their experiment does not readily speak to how potentially mismatching cues are *combined* across distinct levels of linguistic hierarchy online, and whether morphosyntactic inferences are computed in the presence of subsequent disambiguating information. This is different from the current experiment: Here, listeners heard an acoustic cue (the ambiguous schwa), based on which gender and, consequentially, the lexical identity of the target word could be inferred. Crucially, this inference-based lexical preselection could either match or mismatch the identity of the target noun. In contrast to Brown et al. (2012), the subsequent gender cues from adjective and target item in our experiment were always reliable and could, in principle, entirely disambiguate the ambiguous schwa (but importantly, only "in retrospect"). To summarize, our experiment investigates how contextual speech rate (which is an early perceptual, signal-based cue), gender morphology (which is an inferred knowledge-based cue), and lexical information are combined online during spoken language comprehension.

Analysing a time window after the onset of the disambiguating adjective and target noun allowed us to address our third question: How does the system handle incoherence between early perceptual and higher-level linguistic cues when integrating lexical information? By this point in time, participants had already encountered the "unreliable" schwa gender cue ("unreliable" because perception of the ambiguous schwa phoneme should be affected by our rate manipulations), as well as the relatively "reliable" gender cue carried in the adjective and the target word itself. There are three plausible scenarios for how these two cues could be integrated: First, it is possible that the earliest cue completely dominates the later cues as soon as it enters the system. If that were the case, we should observe clear rate effects, and no potential revision based on cues in the target time window. Second, it is possible that participants perceive the first cue as so unreliable that it is immediately overridden as soon as more reliable target cues become available. If that were the case, we should observe no effects of contextual speech rate during the target window. Third, it is possible that both cues are active in the target window to a certain extent. After all, taking all the available information into account would seem to be the best protection against fallibility. If that were the case, the early perceptual cue should remain active in the system for as long as it is relevant for linguistic processing, and we may observe rate effects even after the onset of the disambiguating target information. This is especially interesting given that phoneme-level contextual rate effects have been claimed to be "fragile" (Baese-Berk et al., 2019). As such, our experiment offers novel insights into how the brain infers linguistic cues from the acoustic signal, and how these inferential cues might be combined with information from higher levels of linguistic hierarchy during online sentence comprehension.

2. Methods

Our aim was to test whether and how contextual speech rate influences morphosyntactic and lexical prediction and integration. We used eye-tracking (visual world paradigm) in order to obtain online measures of the influence of contextual rate on the perception of the presence or absence of the morphosyntactic inflectional suffix /-ə/, marking gender on indefinite determiners (feminine *eine* /aɪnə/ vs. neuter *ein* /aɪn/) in German.

2.1 Participants

Native German speakers ($N = 35$, 26 females, $M_{age} = 22$ years) with normal hearing were recruited from the Max Planck Institute (MPI) participant pool, with informed consent as approved by the Ethics Committee of the Social Sciences Department of Radboud University (Project Code: ECSW2014-1003-196). Participants were paid for their participation. We excluded five participants from the analysis due to calibration failures, leaving us with $N = 30$ (23 females, $M_{age} = 23$ years).

2.2 Materials and design

Auditory stimuli consisted of 25 German sentences (e.g., *Schauen Sie jetzt sofort auf ein(e) außergewöhnlich liebe(s) Katze$_{FEM}$/Reh$_{NEU}$*, “Now look at an exceptionally friendly cat/deer”; see Appendix for a complete list of all the stimuli), all sharing the same sentence frame but ending in either a feminine (e.g., *liebe Katze*) or a neuter target reference (*liebes Reh*). Feminine-neuter target pairs were selected that did not have any phonological overlap between the two target nouns (see Appendix). We recorded a female native speaker of German, who was naïve to the purpose of the experiment, reading all sentences with either target reference, but always with the determiner *eine*. Recordings were made in a sound-attenuated booth and digitally sampled at 44,100 Hz on a computer located outside the booth with Audacity software.

For each sentence, the lead-in carrier sentence (*Schauen Sie jetzt sofort auf*) was compressed or expanded in order to yield a fast (66% original duration), a neutral (100% original duration), and a slow (1 / 66% = 150% original duration) syllable rate using PSOLA in Praat (Boersma & Weenink, 2012). Moreover, the duration of the suffix /-ə/ on all determiners *eine* was manipulated. Specifically, 5-step duration continua were created for each recorded *eine* by compressing the word-final schwa using PSOLA in Praat, ranging from perceived absence (40% original duration) to perceived presence (52% original duration) of the schwa phoneme, in steps of 3% (based on piloting). This resulted in a total of 750 unique stimuli (25 sentences x 2 target references x 3 rates x 5 schwa durations).[1]

A categorisation pretest was conducted in order to (1) verify that the duration continua systematically shifted perception from absence to presence of the schwa phoneme; and (2) verify that faster speech rates would bias listeners to explicitly report hearing *eine* (instead of *ein*). Native speakers of German who did not participate in any of the other experiments ($N = 6$, 3 females, $M_{age} = 26$) listened to excerpts (i.e., incomplete sentences) of 250 randomly selected manipulated sentences. Specifically, these excerpts included all the speech up to the disambiguating adjective (e.g., *Schauen Sie jetzt sofort auf ein(e) außergewöhnlich*), thus avoiding biasing influences from the target references on determiner categorisation. Listeners indicated via button press whether they had heard *ein* or *eine*. The categorisation curves (Figure 2) clearly showed that (1) higher steps on the duration continua (i.e., longer schwa) led to more *eine* responses (i.e., fewer *ein* responses); and (2) faster rates (indicated by the different coloured lines in Figure 2) clearly shifted perception towards more *eine* responses. Note that in the eye-tracking experiment, only stimuli from the fast and slow condition were used (no neutral rate condition). Visual stimuli consisted of pictures taken from the MultiPic database (Duñabeitia et al., 2018) presented in 300 x 300 pixel resolution.

[Figure 2]

Figure 2. Categorization curves from the pretest of the proportion of schwa present (i.e., eine) responses as a function of duration continuum step, split for three different contextual speech rates (red: fast rate; green: neutral rate; blue: slow rate). Participants in the pretest only heard short excerpts from the stimulus sentences and indicated whether they heard ein or eine. Longer schwa durations (e.g., step 5) led to more eine responses (i.e., fewer ein responses) and faster speech rate biased listeners to report more eine responses. Error bars represent the standard error of the mean.

In order to minimise the duration of the experiment, participants were randomly allocated to one of two groups: one group was presented with 13 sentences in all possible conditions (13 sentences x 2 target references x 2 rates x 5 duration steps = 260 trials total), the other group with the remaining 12 sentences in all possible conditions (240 trials total). The presentation of the stimuli was randomised in each block, such that all sentences were presented to the participant once before a repetition occurred.

2.3 Procedure

Participants were tested individually in a sound-conditioned booth. They were seated at a distance of approximately 60 cm in front of a 50,8 cm by 28,6 cm screen with a tower-mounted Eyelink 1000 eye-tracking system (SR Research) and listened to stimuli at a comfortable volume through headphones. Stimuli were delivered using Experiment Builder software (SR Research Ltd.). Eye movements were recorded using right pupil-tracking at a sampling rate of 1000 Hz.

Each experimental session started with a nine-point calibration procedure followed by a validation procedure. Participants' task was to listen to the stimuli and click with the computer mouse on one of two pictures corresponding to the two possible sentence-final target references. Note that participants were thus not making any explicit judgment about whether or not they perceived a schwa. In fact, they were ignorant about the intent of the schwa duration and speech rate manipulations. The visual stimuli were presented centred in the left and right halves of the screen. The side of the neuter and female option on the screen was counterbalanced.

On each trial, participants first had to click with the computer mouse on a blue rectangle in the middle of the screen to centre their eye gaze and mouse position. This screen was immediately followed by two pictures. After one second of preview, the auditory stimulus was presented. Participants could only respond by clicking on one of the presented pictures after sound offset. The pictures stayed on the screen until the participant responded by clicking on one of the presented pictures. After an inter-trial interval of one second following the mouse click, the next trial started automatically. Participants first completed a practice session with four trials to become familiarised with the task. Every 80 trials, participants were allowed to take a self-paced break. The experiment took about 35 minutes to complete.

3. Results

Prior to the analyses, blinks and saccades were excluded from the data. We divided the screen into two sections (left and right) and coded fixations on either half as a look toward that particular picture. The eye fixation data were down-sampled to 100 Hz. Participants were very accurate at performing the task: less than 0.2% of the mouse responses were incorrect (*n* = 10). Since the number of incorrect responses was so low, and because we were primarily interested in eye movements prior to and shortly after target onset rather than mouse clicks, we did not exclude any trials from the analyses.

Mixed effects logistic regression models (GLMMs: Generalised Linear Models; Quene & van den Bergh, 2008) with a logistic link function (Jaeger, 2008) as implemented in the MixedModels package (Bates et al., 2015) version 2.1.2+ in Julia version 1.2.0 (Bezanson et al., 2017) evaluated participants' eye fixations. The eye fixation data were evaluated in two time windows: one pre-target time window following the offset of the ambiguous schwa token, and one post-target time window following the onset of the earliest disambiguating target cue. Note that, in cases of a feminine target, the earliest reliably disambiguating cue was the onset of the target noun itself, whereas for a neuter target, the earliest cue was the onset of the morpheme *–s* on the adjective, marking neuter gender.

3.1 Pre-target window

The analysis of the data in the pre-target time window tested whether participants showed an anticipatory target preference – well before the target reference – triggered by the schwa duration in the determiner and the contextual speech rate. The time window of interest was defined as starting from 200 ms after the offset of the ambiguous schwa phoneme, because the offset is the earliest time point at which participants have access to the duration cues on the schwa (note that 200 ms corresponds approximately to the time it takes to launch a saccade; Matin, Shao, & Boff, 1993) and lasting until the onset of the earliest disambiguating cue. For feminine target references, this is the onset of the target word itself; for neuter targets, it is the onset of the morpheme *-s* on the adjective preceding the target word. Figure 3 shows fixation proportions to the feminine picture depending on the context rate (slow vs. fast rate), with the time window of interest shaded grey.

[Figure 3]

Figure 3. Proportion of looks to feminine object across time in fast (red) and slow contexts (blue). Time point 0 marks the offset of the ambiguous schwa phoneme, indicated by the solid vertical line. The dotted vertical line indicates the mean onset of the disambiguating sound: for feminine target references, this is the onset of the target word itself; for neuter targets, it is the onset of the morpheme -s on the adjective preceding the target word. Shown in grey is the area of interest, spanning from 200 ms after schwa offset until the mean onset of the disambiguating cue. Overall, the proportion of looks to the feminine object was higher in fast as opposed to slow contexts. Shading around the coloured lines represent the standard error of the mean.

We predicted that a fast speech rate would bias the perception of the ambiguous determiner *ein[?]* towards *eine* (and away from *ein*) and would trigger more looks to the feminine picture well before the target referent had been heard. Conversely, the slow speech rate would bias perception towards *ein* and, as a consequence, would induce more looks to the neuter picture. Since no phonetic information about the target was available to the listener in the pre-target time window, we analysed participants' looks to just one of the two objects (the feminine object, instead of looks to the target), coded binomially.

A generalised linear mixed model with a logistic linking function tested the binomial looks to the feminine picture (1 = yes, 0 = no) for fixed effects of Rate (categorical predictor with two levels: fast coded as +0.5, slow as -0.5), Time (continuous predictor; z-scored around the mean within the analysis window), Step (continuous predictor; centred: schwa duration continuum Step 1 coded as -2, Step 3 as 0, Step 5 as 2), and their interactions. Additionally, the model included a fixed effect of Lag, capturing the binomial looks to the feminine picture at the previous sample (1 = yes, 0 = no). The Lag predictor addresses the autocorrelated nature of eye gaze data arising from the way that visual world eye-tracking data are collected: fixations last longer than a single sample, so the probability of a fixation at sample $s+1$ is conditional upon whether there had been a fixation at sample s (cf. Cho, Brown-Schmidt, & Lee, 2018). The random effects structure contained random intercepts for Participants and Items and by-participant and by-item slopes for all fixed factors including Lag (but not their interactions).

The model revealed a significant effect of Rate (β = 0.114, SE = 0.047, z = 2.481, p = 0.013), demonstrating that upon hearing an ambiguous phoneme, participants were more likely to look at the feminine object during trials that included a fast context rate. Crucially, this happened before the onset of any further disambiguating cues. We also found a significant interaction between Time and Step (β = 0.021, SE = 0.007, z = 3.128, p = 0.002), indicating that higher continuum steps led to an increasingly higher proportion of looks to the feminine object as time passed. Finally – and unsurprisingly – , the model revealed a significant main effect of Lag, indicating that looks to the feminine object were, indeed, dependent on the gaze at the previous sample ($\beta = 8.215$, SE = 0.089, z = 91.678, $p < 0.001$). Overall, these results support our hypotheses: Participants were more likely to look at the picture corresponding to the feminine object in the fast rate, thus indicating that they were more likely to have perceived a schwa phoneme in the fast as opposed to the slow context, and that they used that percept as a morphological gender cue towards the target picture.

3.2 Post-target window

The analysis of the data in the post-target time window tested whether the effects of contextual rate and schwa duration manipulations persisted even after the perception of disambiguating phonological cues (i.e., after target onset). The time window of interest was defined as starting from 200 ms after the onset of the earliest disambiguating cue (target word onset for feminine, *-s* morpheme onset for neuter targets) and lasting until 200 ms after the offset of the target word's initial syllable. As noted above, there was no phonological overlap between target and competitor images, so the earliest target-specific acoustic cues can, in principle, entirely disambiguate between the two. Evidence of this can be seen in Figure 4, where we observe preferential looks towards the target picture well before the offset of the first syllable of the target.

[Figure 4]

Figure 4. Proportion of looks to target object across time for feminine targets (solid) and neuter targets (dashed) in fast (red) and slow contexts (blue). The feminine-fast (solid red line) and neuter-slow (dashed blue line) conditions represent the Congruent conditions; the feminine-slow and neuter-fast conditions represent the Incongruent conditions. Time point 0 marks the onset of the earliest disambiguating cue (onset of the target word for feminine targets, morpheme -s on the preceding adjective for neuter targets), indicated by the vertical solid line. The vertical dotted line indicates the mean offset of the first target word syllable. Shown in grey is the area of interest, spanning from 200 ms after onset of the disambiguating cue until 200 ms after the mean offset of the initial target word syllable. Shading around the coloured lines represent the standard error of the mean.

We had crossed the factors rate and target gender. According to our predictions (and as shown in the pretest), an ambiguous /-ə/ token presented in a *fast* context is more likely to be perceived as *present*. In terms of our experimental manipulation, the perceived presence of a schwa phoneme corresponds to the perception of the determiner *eine*, marking feminine gender. Fast context rates should therefore bias participants' looking preference towards the picture corresponding to the feminine object. We therefore refer to trials with a feminine target presented in a fast context sentence as *rate-gender congruent trials*. Similarly, an ambiguous /-ə/ token presented in a *slow* context is more likely to be perceived as *absent*, thus corresponding to the perception of the neuter determiner *ein* and eliciting more looks towards the picture corresponding to the neuter object. Therefore, trials with a neuter target presented in a slow context sentence are also referred to as *rate-gender congruent trials*. Conversely, *feminine+slow* and *neuter+fast* trials are referred to as *rate-gender incongruent*. The use of this congruency coding allowed us to specifically test for potentially facilitating effects of *congruent* contextual speech rate on target looks, independent of the speech rate in a given trial. As can be seen in Figure 4, participants seemed to be faster to look at the correct target picture in congruent as opposed to incongruent trials.

A GLMM with a logistic linking function tested the binomial looks to the target picture (1 = yes, 0 = no) for fixed effects of Congruency (categorical predictor with two levels: congruent coded as +0.5; incongruent as -0.5), Step (continuous predictor; centred: schwa duration continuum Step 1 coded as -2, Step 3 as 0, Step 5 as 2), and Time (continuous predictor; z-scored around the mean within the analysis window), and all their interactions. Again, we also included a Lag predictor (categorical predictor coding looks to the target picture at the previous sample: 1 = yes, 0 = no) in order to alleviate the autocorrelation problem. The random effects structure contained random intercepts for Participants and Items and by-participant and by-item random slopes for all fixed factors including Lag (but not their interactions).

The model revealed a significant effect of Time (β = 1.635, SE = 0.120, z = 13.611, $p < 0.001$), indicating, unsurprisingly, that participants increasingly looked at the target picture as time passed. Crucially, a significant effect of Congruency was found (β = 0.124, SE = 0.060, z = 2.081, $p = 0.038$), indicating that participants showed more looks to the target referent if the preceding morphological cue, inferred from the perceived presence or absence of the schwa phoneme based on contextual speech rate, was "congruent" with the target gender (e.g., fast with feminine targets; slow with neuter targets). No effect of Step could be established (β = -0.024, SE = 0.020, z = -1.193, $p = 0.233$). This is not surprising, considering that low Steps would have biased participants towards perceiving a schwa as *not* being present (thus leading to a neuter interpretation), and high Steps would have biased participants toward perceiving a schwa as *being* present (thus leading to a feminine interpretation); since half of the targets were neuter and the other half were feminine, any biasing effect of Step simply averages out between the two target genders.

Moreover, several interactions were observed. An interaction between Congruency and Time (β = -0.245, SE = 0.029, z = -8.330, $p < 0.001$) indicated that the beneficial effect of a congruent speech rate diminished with time. However, a positive three-way interaction (β = 0.109, SE = 0.020, z = 5.324, $p < 0.001$) between Congruency, Step and Time indicated that this only held for the lower continuum steps. The model also found an interaction between Congruency and Step (β = 0.103, SE = 0.019, z = 5.324, $p < 0.001$), indicating that the effect of Congruency was smaller for lower continuum steps (i.e., shorter schwa durations). This may be interpreted in light of the pretest: The rate effect was smaller at lower continuum steps (cf. Figure 2), and as such the effect of congruency would also be expected to be smaller. Finally, we found an interaction between Time and Step (β = 0.046, SE = 0.011, z = 4.358, $p < 0.001$);

although we currently lack an interpretation for this interaction, note that the estimate is very small. Finally – and again as expected – the model revealed a significant main effect of Lag, indicating that looks to the target object were, indeed, dependent on the gaze at the previous sample ($\beta = 6.971$, SE = 0.070, z = 99.391, $p < 0.001$).

4. Discussion

The aim of the current study was to investigate three main questions. First, we asked whether we could observe early perceptual cues being rapidly used online in order to infer higher-level linguistic cues, even in the presence of subsequent disambiguating information. Second, we asked whether these inferential cues that were based on perceptual cues are deployed to make predictions about upcoming linguistic information. Third, we asked how the language comprehension system handles incoherence between early perceptual and higher-level linguistic cues when integrating lexical information. We addressed these questions by experimentally inducing contextual rate normalisation effects on the phoneme /-ə/, which can act as a morphosyntactic gender cue on indefinite determiners in German. In the following, we will discuss our results in light of these three questions.

4.1 Contextual speech rate is rapidly used as a cue for speech processing

We found evidence for contextual speech rate acting as an early and robust cue for speech comprehension. Listeners' perception of the morpheme /-ə/ in German was significantly influenced by the rate of the preceding context. We observed these rate normalisation effects immediately after the presentation of the ambiguous schwa token (200 ms after schwa offset), and well before any acoustic information about the target referent itself was available to the listeners. These results support previous accounts of rate normalisation effects arising during early stages of lexical processing and influencing phoneme perception almost immediately (Bosker, 2017a; Newman & Sawusch, 2009; Reinisch & Sjerps, 2013; Toscano & McMurray, 2015; Kaufeld et al., in press; Maslowski et al., 2019b).

Our findings are novel in two ways. First, to our knowledge, previous eye-tracking studies on contextual rate normalisation have mostly investigated minimal word pairs (e.g., *tak* vs. *taak* (Kaufeld et al., in press; Reinisch & Sjerps, 2013); *tear* vs. *terror* (Baese-Berk et al., 2019); *eens speer* vs. *een speer* (Reinisch, Jesse, & McQueen, 2011), where the interpretation of the ambiguous phoneme had implications on a lexical level, but did not affect further linguistic processing on the sentence level (although see Brown et al., 2012). In contrast, the rate manipulation in the current experiment affected the perception of a purely morphosyntactic minimal pair (*ein* vs. *eine*). Here, we show for the first time how contextual speech rate – an acoustic, signal-based cue – interacts online with subsequent gender information from a lexical, knowledge-based cue, which occurs on a higher level of the linguistic hierarchy and was potentially conflicting with the earlier cue. As such, this is the first eye-tracking experiment to our knowledge where the rate manipulation carried implications for further inference-based morphosyntactic prediction and integration of subsequent lexical material. Second, previous research has mostly used experimental tasks that involved explicit identification or categorisation of the ambiguous word. In contrast to that, our design allowed us to tap perception of the ambiguous determiner *ein?*, crucially without explicitly asking participants for a categorisation decision between *ein* and *eine*. This contrasts with earlier eye-tracking studies of rate normalisation (e.g., Reinisch & Sjerps, 2013; Toscano & McMurray, 2012, 2015; Kaufeld et al., in press) where participants did make explicit categorisation decisions about the ambiguous target sounds under study. Notably, this is also different from the experiment reported by Brown et al.

(2012), where participants decided between singular or plural targets and thus made explicit judgments about the informational content of the phoneme affected by the rate manipulation. As such, our results suggest that rate normalisation operates automatically, even when attention is not drawn to the ambiguous target sounds tested. This corroborates recent findings from Maslowski et al. (2019b), who showed evidence that listeners normalize for speech rate even without an explicit recognition task (using repetition priming). In light of these two aspects, our findings demonstrate that 1) rate normalisation affects a large set of duration-cued distinctions, including morphosyntactic minimal pairs, and 2) rate normalisation impacts incremental spoken language processing, even when the task does not require participants to make explicit judgments. As such, rate normalisation observed in lab-based psycholinguistic experiments appears to be a perceptual process that likely also contributes to the comprehension of natural and spontaneous conversation.

4.2 Inferences that were made based on perceptual cues can be used as higher-level cues to make predictions about upcoming referents

As stated above, our experiment went beyond mere phonemic or lexical identification: The indefinite determiner containing the ambiguous schwa token was the first cue towards the gender of the target picture, so it was a crucial building block for subsequent steps of language processing. Our eye gaze analysis in a time window after the ambiguous schwa token showed that participants not only immediately made use of contextual information upon perceiving the ambiguous token, but also rapidly used that information to draw inferences about the gender of the target referent. This was reflected in participants looking more towards the picture that corresponded to the gender that the rate manipulation biased them towards.

Our experiment contributes to the current debate around prediction during language comprehension (cf. Huettig, 2015; Nieuwland et al., 2018). Several studies have found effects of anticipatory language processing based on gender information carried in the determiner (e.g., Szewczyk & Schriefers, 2013; Van Berkum et al., 2005; Wicha, Bates, et al., 2003; Wicha et al., 2004; Wicha, Moreno, et al., 2003), while others have failed to replicate these findings (Kochari & Flecken, 2019; Guerra et al., 2018). Why is it that we find evidence for anticipatory language processing in our current experiment, while others did not? One reason might be that we provided participants with the same fixed sentence frame on every trial, making the *ein/eine* distinction relatively informative – possibly more so than it would be in more naturalistic settings. Moreover, language comprehension occurred within a very small referential "world" in our experiment: Participants were presented with two pictures at a time, thus limiting their choices for possible predictions considerably. Presumably, these two factors facilitated the predictive processing observed. Nevertheless, the fact that rate normalisation induces the kind of predictive behaviour that we observe with our paradigm is strong evidence for the utility of contextual rate cues in speech processing.

As stated earlier (see section 2.3), prediction is a possibility, but not a necessity, for language comprehension within a cue integration framework. We therefore do not take our findings as evidence in favour of, or against anticipatory language processing, per se; rather, we believe that our results can be seen as step towards a more comprehensive account of language processing where predictions *can* be part of the processing architecture.

4.3 Early perceptual cues remain active in the speech and language comprehension system during subsequent processing

Even after hearing the disambiguating beginning of the target referent, participants were significantly slower to look at the target object in *rate-gender incongruent* trials (i.e., in trials where the actual target gender did not match the gender corresponding to the schwa perception induced by the preceding context rate manipulation). We believe that this finding – a robust effect of a low-level perceptual cue, even in the presence of the reliably unambiguous first syllable of the target word – indicates that the early perceptual cue does, indeed, remain active in the system, until it can (or cannot) be integrated with additional incoming information.

These observations are in line with previous behavioural studies (Heffner et al., 2015; Morrill et al., 2015), where rate effects also persisted even in the presence of constraining higher-level linguistic information. Crucially, using the visual-world paradigm allowed us to measure responses to the rate manipulation without asking for explicit categorisation of *ein* vs. *eine,* so in contrast to previous studies, no task-driven attention was drawn to the ambiguous sounds. Taken together, these findings suggest that phoneme-level rate effects are not "fragile", as has previously been suggested, but rather that they are robust and persist even in the presence of higher-level linguistic (in our case lexical) information. Interestingly, results reported by Morrill et al. (2015) and Kaufeld et al. (in press) suggest that listeners are flexible in the way that they weigh specific cues, depending on the context and listening situation. Models of cue integration can accommodate these results, given that cue weights can be updated dynamically depending on the cue's reliability within a given situation.

Our observations also speak to recent findings by Gwilliams et al. (2018). They reported online MEG evidence showing that sensitivity to phoneme ambiguity occurs at the earliest sensory stages of speech processing, and that this sensitivity to ambiguity, along with other fine-grained acoustic features such as VOT, appeared to be maintained throughout later processing stages, even as further lexical information entered the system. The authors suggest that this reflects a reassessment of the ambiguous speech sound as additional input is being perceived. We believe that these findings can also be explained within a cue integration architecture: The early perceptual cue remains active for as long as it is relevant for linguistic processing, and its validity and reliability are "reassessed" incrementally as part of sensory resampling, as it is integrated with cues from higher levels of linguistic processing.

Our experiment is not the first to examine contextual speech rate as an early perceptual cue within a cue integration framework. Toscano and McMurray (2012, 2015) have argued that contextual speech rate can modulate the uptake of other phonological cues, such as VOT. They elegantly explain this within the C-CuRE framework: expected values are established based on contextual speech rate, and new cues are computed relative to those expectations. In fact, Toscano and McMurray (2015) suggest that adjusting these expectations can be explained within C-CuRE as a form of predictive coding, and they point out that cue integration models of speech perception have to be linked to lexical processes. Their observations thus fit seamlessly into a more general framework of cue integration for language processing as suggested by Martin (2016), where the system makes use of all relevant pieces of information across different levels of linguistic hierarchies in order to reduce fallibility.

Based on our findings, we can formulate new questions for future research. For example, an iterative model of cue integration would suggest that lower-level perceptual ambiguity would carry through to even higher levels of linguistic processing that go beyond morphosyntax. Future experiments could therefore investigate whether rate normalisation effects induced by contextual speech rate also affect semantic prediction and integration. If so, do early perceptual cues even remain active within a larger discourse? It seems plausible that there would be at least some temporal limit regarding how long early ambiguous cues remain active in the system. If so, it would be desirable to test where that cut-off point might be, or whether it can be dynamically adjusted depending on the reliability of a specific cue in a given situation.

Our experiment investigated two cues, specifically: contextual speech rate and grammatical gender. As Martin (2016) and others have pointed out, one of the hardest definitions to provide within a cue integration framework is what can constitute a cue. Future experiments are thus needed in order to determine an inventory of psycholinguistic cues and examine which other (lower- and higher-level, knowledge- and signal-based) pieces of information the brain draws on to arrive at robust linguistic units and structures (cf. Kaufeld et al., in press).

Finally, with regard to our third question, it might be interesting to investigate in more detail *why* it took participants longer to look at the target picture in *rate-gender incongruent* trials, that is, which sub-mechanisms of cue integration and/or oculomotor control might have caused this delay. One possible explanation would be integration difficulty of the second cue in the presence of the earlier, incongruent cue. This integration difficulty could arise from participants generally taking longer to integrate the mismatching cue, but it is also possible that participants attempted the integration process multiple times and therefore took longer to converge on the target. Another possible explanation might be a “spill-over” effect, where participants were slower to look at the target in incongruent trials because of the additional time it took them to first shift their gaze, either by cancelling a previously planned saccade, or by initiating an entirely new saccade (see Altmann, 2011, for a general discussion of language-mediated eye movements). Though this was not the focus of our current experiment, investigating the subroutines at play during the integration of incongruent cues in more detail may be an interesting objective for further research.

Taken together, our results show that contextual rate effects rapidly influence not only lexical processing, but also subsequent morphosyntactic prediction and integration. Linguistic models of cue integration offer a promising step towards a mechanistic explanation for how the brain accomplishes the task of inferring complex meaning from a noisy acoustic signal by operationalizing both lower-level, perceptual and higher-level, linguistic cues.

5. Acknowledgements

We would like to thank Nina Mainz for lending her voice for the stimuli recordings, Phillip M. Alday for advice on the statistical analyses, and Joe Rodd for help with Figure 1. This research was supported by funding from the Max Planck Research Group "Language and Computation in Neural Systems" and from the Netherlands Organization for Scientific Research (grant 016.Vidi.188.029) to AEM.

6. References

Altmann, G. T. M. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. Acta Psychologica, 137(2), 190–200. https://doi.org/10.1016/j.actpsy.2010.09.009

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition, 73(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Apfelbaum, K. S., Bullock-Rest, N., Rhone, A. E., Jongman, A., & McMurray, B. (2014). Contingent categorisation in speech perception. Language, Cognition and Neuroscience, 29(9), 1070–1082. https://doi.org/10.1080/01690965.2013.824995

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L., & Banzina, E. (2019). Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables. *Attention, Perception, & Psychophysics*, *81*(2), 571-589.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. *Mechanisms of language acquisition*, 157-193.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. doi: 10.18637/jss.v067.i01

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. SIAM Review, 59, 65-98.

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer (Version 5) [52]. Retrieved from http://www.praat.org/

Bölte, J., & Connine, C. M. (2004). Grammatical gender in spoken word recognition in German. Perception & Psychophysics, 66(6), 1018–1032. https://doi.org/10.3758/BF03194992

Bosker, H. R. (2017a). Accounting for rate-dependent category boundary shifts in speech perception. Attention, Perception, & Psychophysics, 79(1), 333–343. https://doi.org/10.3758/s13414-016-1206-4

Bosker, H. R. (2017b). How our own speech rate influences our perception of others. Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(8), 1225–1238. https://doi.org/10.1037/xlm0000381

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. Language, Cognition and Neuroscience, 33(8), 955–967. https://doi.org/10.1080/23273798.2018.1439179

Bosker, H. R., & Reinisch, E. (2017). Foreign Languages Sound Fast: Evidence from Implicit Rate Normalization. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01063

Bosker, H. R., Reinisch, E., & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. Journal of Memory and Language, 94, 166–176. https://doi.org/10.1016/j.jml.2016.12.002

Brown, M., Dilley, L. C., & Tanenhaus, M. K. (2012). Real-time expectations based on context speech rate can cause words to appear or disappear. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, No. 34).

Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. Science, 264(5163), 1333–1336. https://doi.org/10.1126/science.8191289

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. Nature Reviews Neuroscience, 13(1), 51–62. https://doi.org/10.1038/nrn3136

Cho, S. J., Brown-Schmidt, S., & Lee, W. Y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: an application to intensive binary time series eye-tracking data. *Psychometrika*, *83*(3), 751-771.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(03), 181–204. https://doi.org/10.1017/S0140525X12000477

Connine, C. M., & Clifton, C. (1987). Interactive Use of Lexical Information in Speech Perception. Journal of Experimental Psychology: Human Perception and Performance, 13(2), 291–299.

Dell, G. S., & Chang, F. (2013). The P-chain: relating sentence production and its disorders to comprehension and acquisition. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1634), 20120394–20120394. https://doi.org/10.1098/rstb.2012.0394

Dilley, L. C., & Pitt, M. A. (2010). Altering Context Speech Rate Can Cause Words to Appear or Disappear. Psychological Science, 21(11), 1664–1670. https://doi.org/10.1177/0956797610384743

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. Quarterly Journal of Experimental Psychology, 71(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. Trends in Cognitive Sciences, 8(4), 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. Journal of Memory and Language, 41(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., McLennan, D. B., Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. Psychophysiology, 39(2), 133–146. https://doi.org/10.1111/1469-8986.3920133

Fetsch, C. R., DeAngelis, G. C., & Angelaki, D. E. (2013). Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nature Reviews Neuroscience, 14(6), 429–442. https://doi.org/10.1038/nrn3503

Fox, R. A. (1984). Effect of Lexical Status on Phonetic Categorization. Journal of Experimental Psychology: Human Perception and Performance, 10(4), 526–540.

Friederici, A. D., & Jacobsen, T. (1999). Processing Grammatical Gender During Language Comprehension. Journal of Psycholinguistic Research, 28(5), 467–484.

Ganong, W. F. (1980). Phonetic Categorization in Auditory Word Perception. Journal of Experimental Psychology: Human Perception and Performance, 6(1), 110–125.

Gordon, P. C. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech. Perception & Psychophysics, 43(2), 137–146. https://doi.org/10.3758/BF03214191

Guerra, E., B. Nicenboim, and A. V. Helo (2018). “A crack in the crystal ball: Evidence against pre-activation of gender features in sentence comprehension”. *Poster presented at the AMLaP conference (Architectures and Mechanisms for Language Processing)*. Berlin, Germany.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In Spoken Word Recognition, the Future Predicts the Past. The Journal of Neuroscience, 38(35), 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018

Heffner, C. C., Newman, R. S., Dilley, L. C., & Idsardi, W. J. (2015). Age-related differences in speech rate perception do not necessarily entail age-related differences in speech rate use. *Journal of Speech, Language, and Hearing Research*, *58*(4), 1341-1349.

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, & Psychophysics*, *79*(3), 964-988.

Hillert, D., & Bates, E. (1994). Morphological Constraints on Lexical Access: Gender Priming in German. La Jolla: Center for Research in Language, University of California, San Diego.

Huettig, F. (2015). Four central questions about prediction in language processing. Brain Research, 1626, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. Language, Cognition and Neuroscience, 31(1), 19–31. https://doi.org/10.1080/23273798.2015.1072223

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of Syntactic and Semantic Information in Predictive Processing: Cross-Linguistic Evidence from German and English. Journal of Psycholinguistic Research, 32(1), 37–55.

Kaufeld, G., Ravenschlag, A., Meyer, A. S., Martin, A. E., & Bosker, H. R. (in press). Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition. https://doi.org/10.1037/xlm0000744

Kochari, A. R., & Flecken, M. (2019). Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, *34*(2), 239-253.

Landy, M. S., Banks, M. S., & Knill, D. C. (2011). Ideal-Observer Models of Cue Integration. In J. Trommershäuser, K. Kording, & M. S. Landy (Eds.), Sensory Cue Integration (pp. 5–29). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387247.003.0001

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. Journal of Memory and Language, 61(3), 326–338. https://doi.org/10.1016/j.jml.2009.06.004

Martin, A. E. (2016). Language Processing as Cue Integration: Grounding the Psychology of Language in Perception and Neurophysiology. Frontiers in Psychology, 7. https://doi.org/10.3389/fpsyg.2016.00120

Martin, A. E., Monahan, P. J., & Samuel, A. G. (2017). Prediction of Agreement and Phonetic Overlap Shape Sublexical Identification. Language and Speech, 60(3), 356–376. https://doi.org/10.1177/0023830916650714

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2018). Listening to yourself is special: Evidence from global speech rate tracking. PLoS One, 13(9): e0203571. doi:10.1371/journal.pone.0203571.

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019a). How the tracking of habitual rate influences speech perception. Journal of Experimental Psychology: Learning, Memory, and Cognition 45(1), 128.

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019b). Listeners normalize speech for contextual speech rate even without an explicit recognition task. The Journal of the Acoustical Society of America.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. Perception & Psychophysics, 53(4), 372–380. https://doi.org/10.3758/BF03206780

Mattys, S. L., Melhorn, J. F., & White, L. (2007). Effects of syntactic expectations on speech segmentation. Journal of Experimental Psychology: Human Perception and Performance, 33(4), 960.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General*, *134*(4), 477.

McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In Where do features come from (pp. 197–236).

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review, 118(2), 219–246. https://doi.org/10.1037/a0022325

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of/b/and/w. *The Journal of the Acoustical Society of America*, *73*(5), 1751-1755.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. Journal of Experimental Psychology: Human Perception and Performance, 14(3), 369.

Mitterer, H. (2018). The singleton-geminate distinction can be rate dependent: Evidence from Maltese. Laboratory Phonology: Journal of the Association for Laboratory Phonology, 9(1), 6. https://doi.org/10.5334/labphon.66

Morrill, T., Baese-Berk, M., Heffner, C., & Dilley, L. (2015). Interactions between distal speech rate, linguistic knowledge, and speech environment. *Psychonomic bulletin & review*, *22*(5), 1451-1457.

Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. Journal of Phonetics, 37(1), 46–65. https://doi.org/10.1016/j.wocn.2008.09.001

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. ELife, 7. https://doi.org/10.7554/eLife.33468

Oden, G. C., & Massaro, D. W. (1978). Integration of Featural Information in Speech Perception. Psychological Review, 85(3), 172–191.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? Trends in Cognitive Sciences, 11(3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002

Pickett, J. M., & Decker, L. R. (1960). Time Factors in Perception of A Double Consonant. Language and Speech, 3(1), 11–17. https://doi.org/10.1177/002383096000300103

Pitt, M. A., & Samuel, A. G. (1993). An Empirical and Meta-Analytic Evaluation of the Phoneme Identification Task. Journal of Experimental Psychology: Human Perception and Performance, 19(4), 699–725.

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. https://doi.org/10.3758/s13414-015-0981-7

Quene, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling.

Rabagliati, H., & Bemis, D. K. (2013). Prediction is no panacea: The key to language is in the unexpected. Behavioral and Brain Sciences, 36(4), 372–373. https://doi.org/10.1017/S0140525X12002671

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. Journal of Experimental Psychology: Human Perception and Performance, 37(3), 978–996. https://doi.org/10.1037/a0021923

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. Journal of Phonetics, 41(2), 101–116. https://doi.org/10.1016/j.wocn.2013.01.002

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. https://doi.org/10.1037/a0026786

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. Neuropsychologia, 51(3), 437–447. https://doi.org/10.1016/j.neuropsychologia.2012.12.002

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. Perception & Psychophysics, 62(2), 285-300.

Sawusch, J. R., & Pisoni, D. B. (1974). On the identification of place and voicing features in synthetic stop consonants, Journal of Phonetics, 2, 181-194.

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. Journal of Memory and Language, 68(4), 297–314. https://doi.org/10.1016/j.jml.2012.12.002

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. Attention, Perception, & Psychophysics, 74(6), 1284–1301. https://doi.org/10.3758/s13414-012-0306-z

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: effects of sentential rate and vowel length on voicing judgments. Language, Cognition and Neuroscience, 30(5), 529–543. https://doi.org/10.1080/23273798.2014.946427

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. https://doi.org/10.1037/0096-1523.27.5.1057

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and

Reading Times. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. Perception & Psychophysics, 67(6), 939–950. https://doi.org/10.3758/BF03193621

Wicha, N. Y. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences. Neuroscience Letters, 346(3), 165–168. https://doi.org/10.1016/S0304-3940(03)00599-8

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2003). Expecting Gender: An Event Related Brain Potential Study on the Role of Grammatical Gender in Comprehending a Line Drawing Within a Written Sentence in Spanish. Cortex, 39(3), 483–508. https://doi.org/10.1016/S0010-9452(08)70260-0

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating Words and Their Gender: An Event-related Brain Potential Study of Semantic Integration, Gender Expectancy, and Gender Agreement in Spanish Sentence Reading. Journal of Cognitive Neuroscience, 16(7), 1272–1288. https://doi.org/10.1162/0898929041920487

[1] Note that a distinction is commonly made between *distal* and *proximal* speech rate manipulations (see Heffner, Newman, and Idsardi (2017) for an in-depth discussion of this distinction), where *proximal* context refers to the context directly adjacent to the ambiguous region of interest, whereas *distal* context refers to linguistic material that is further away (i.e., non-adjacent from the ambiguous region of interest. In the current experiment, we are manipulating context that is not directly adjacent to the ambiguous schwa phoneme. That is, the syllable *ein-* intervened between the rate-manipulated context and the ambiguous schwa phoneme; as such, our rate manipulation can be considered *distal.*