Chapter 18:

Phonetics and eye-tracking

Eva Reinisch and Holger Mitterer

18.0. Abstract

Eye-tracking has proven to be a fruitful method to investigate how listeners process spoken language in real-time. For efficient comprehension of the incoming message, listeners have to continuously evaluate the incoming acoustic signal and map it onto prelexical and lexical representations. Eye-tracking makes use of listeners' behaviour to spontaneously fixate on visual referents to spoken input allowing researchers to track speech perception online in a closely time-locked fashion. Importantly, fixations on referents are triggered even by a partial match between the acoustic signal and the referent label. As a consequence, fixations can be taken as a measure of which words listeners temporarily (and unconsciously) consider a possible target as the speech signal unfolds. By comparing fixations on competitors with specific phonetic properties researchers can assess which properties are the most relevant for word recognition and when during word recognition they are most important. Eye-tracking hence allows us to track lexical competition online while asking participants to perform a relatively natural task, namely finding visual referents to what they are hearing.

This chapter highlights the contribution of eye-tracking to our understanding of various classical issues in phonetics about the uptake of segmental and suprasegmental information during speech processing, as well as the role of gaze during speech perception. The review introduces the visual-world paradigm which is the most relevant paradigm for phonetic research and shows how variations of this paradigm can be used to investigate the timing of cue uptake, how speech processing is influenced by phonetic context, how word recognition is affected by connected-speech processes, the use of word-level prosody such as lexical stress, and the role of intonation for reference resolution and sentence comprehension. Importantly, since the eye-tracking record is continuous it allows us to distinguish early perceptual processes from post-perceptual processes. For instance, it is possible to investigate how listeners revise their percept when later coming information is incompatible with earlier information. The chapter also provides a brief note on the most important issues to be considered in teaching and using the method including comments on data processing, data analyses, and interpretation, as well as suggestions for how to implement eye-tracking experiments.

18.1. Introduction

One of the basic issues in speech perception is that the acoustic signal unfolds over time. Sounds are not discrete objects like letters, and spoken words are not separated by silence like written words are separated by blank spaces. Nevertheless, human listeners are surprisingly good at understanding speech, that is, at recognizing discrete words in the continuous signal and putting them together to form a message. A challenge for phonetic and psycholinguistic research has been to understand how listeners are able to do so. Eye-tracking has proven a useful measure to address this issue. It has become increasingly influential in psycholinguistics in general but also in phonetic research because eye movements can provide crucial information about phonetic processing. Participants are typically presented with a visual scene (i.e., "visual-world paradigm")—mostly on a computer monitor—while listening to speech. Where and when they look at a given object is taken to reflect their current interpretation of the speech input. In comparison to other methods, eye-tracking has the advantage of being "online" (like event-related potentials, ERPs), relatively straightforward to interpret (similarly to phonetic identification tasks), but is not strongly influenced by task specific processes (a drawback in many priming studies). It also approximates speech processing outside the lab, as the task in an eye-tracking study is not that different from normal, embodied language processing for requests such as "Could you pass me the salt, please?".

Due to its online nature, eye-tracking can be used to infer how listeners interpret a word or a sentence before all of it is heard (see Figure 18.1 for how eye-tracking data reveal the immediate use of information that is provided as the sentence unfolds). It is hence often used to investigate issues that previously were addressed in studies using gating and fragment priming tasks. For instance, Shockey (2003) reports a study in which participants hear more and more of a spontaneously uttered sentence and have to write down what they heard. Obviously, there is a strong offline component to this task, so that participants may write words that were not actually considered during online speech processing. Eye-tracking has the potential to show which lexical candidates are considered *as they are heard* (see e.g., Brouwer, Mitterer, & Huettig, 2012, for addressing a similar question to Shockley, 2003, using eye-tracking). The eye-tracking method also has the advantage of being relatively straightforward to interpret: When participants look at a picture of an apple with an above chance level, it is reasonable to assume that the input causes them to think about apples, be it, because they activate the word *apple* in their mental lexicon, or because the input is related to apples (see Huettig, Rommers, & Meyer, 2011, for examples of what may trigger eye movements to a picture). This inference is often more straightforward than interpreting ERP components.

18.2. Historical Overview

The studies discussed in this chapter all deal with an experimental design that has become known as the visual world paradigm (VWP). In this paradigm, listeners typically view arrays of pictures or printed words—or in some cases a natural scene such as an array of objects on a table—while listening to speech. Gaze position is recorded and analysed in a time-locked fashion with the acoustic input. Cooper (1974) was the first to show that listeners spontaneously direct their gaze at visual referents to spoken input that are related in phonological form or meaning. For example, upon hearing words such as "lion" or "Africa", listeners are more likely to direct their gaze to the picture of a lion than to other unrelated pictures. Gaze position also provides insights into what types of speech input listeners anticipate. Indeed Cooper (1974) noted that upon hearing the phrase "lion and ..." listeners may direct their gaze away from the lion and towards a zebra located in another part of the display - before the word "zebra" or any other word has been heard.

Interestingly, it was only some 20 years after the pioneering study by Cooper (1974) that the paradigm became more popular for research on language processing. After a seminal study in *Science* on syntactic processing (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), the paradigm gained impact in psycholinguistic and phonetic research. Progress in technologies and computing power may have contributed to this advance over the last two decades, making the setup and especially analyses of data more user-friendly (see *Best Practice*).

For the study of phonetic processing, the first influential study after Cooper (1974) was Allopenna, Magnuson and Tanenhaus (1998) that demonstrated the timecourse of the uptake of segmental information. Their experimental design has become the basis for many if not most studies on phonetic processing. They asked participants to manipulate one of four objects on a screen (e.g., "Pick up the *beaker*; now put it below the diamond"). Visual referents included a picture of the target (*beaker*), a phonetic cohort competitor that overlapped at word onset (*beetle*), a rhyme competitor that overlapped at word offset (*speaker*) and a phonetically and semantically unrelated object (*carriage*). Eye-tracking data showed that the target and cohort competitor were fixated on from target word onset (as they both matched the acoustic input) with competitor fixations starting to decrease as the target became acoustically distinct. Importantly, the rhyme competitor, despite its mismatch at target onset was also fixated on more than the unrelated referent. The timecourse of rhyme competitor fixations was somewhat delayed relative to target and cohort competitor fixations, just as the phonetic overlap between competitor and target is delayed with rhyme competitors in comparison to cohort competitors. These results were taken as evidence that as the acoustic speech signal unfolds, listeners consider all words that temporarily match the acoustic input until one word has gained sufficient support to be the one that listeners tend to identify. Note that this "lexical

competition", that is, the activation of competitor words along with the target lasts beyond the point at which the target becomes phonetically distinct. This supports the idea that spoken word recognition is a probabilistic process. Allopenna et al. (1998) also demonstrated that eye-tracking provides a more sensitive measure of the word recognition process than other methods. They additionally performed a gating experiment, in which participants were presented with successively longer parts of the target words - the same words they had used for eye-tracking. Both methods revealed competition at word onset, but the gating experiment failed to show the effects for rhyme competitors. In addition, the activation patterns of the eye-tracking study, but not the gating study, matched predictions by a computational model of spoken-word recognition (TRACE; McClelland & Elman, 1986).

18.3. Critical Issues

While eye-tracking systems differ in their technical details (see *Best Practice* section), they all aim at measuring the direction of gaze, by sampling eye fixations at rates between 30 and 2000 Hz. Eye-tracking studies are sometimes referred to as "eye-movement studies", but what is typically analysed are not the eye movements themselves. Gaze position is always slightly changing (due to the so-called Nystagmus that prevents neural fatigue), but can be aggregated over time into saccades and fixations. During fixations, eye position is more or less constant, whereas saccades describe fast movements from one position to another. During saccades, neural transduction is inhibited. The parsing of the raw data into these types of events is often achieved by pre-processing algorithms that come with a given eyetracker. Most eye-tracking studies focus on fixation positions, but it is also common to combine a saccade with the following fixation into "looks" (McMurray, Clayards, Tanenhaus, & Aslin, 2008).

As noted earlier, the main advantage of eye-tracking over other methods is that it allows the measurement of speech recognition over time. The measure is closely time locked to the acoustic input with a constant delay of about 150-200ms, which is an estimate how long it takes to plan and execute an eye movement. Eye-tracking is non-disruptive, that is, depending on the exact equipment (and with the exception of possibly being asked to avoid excessive head movements) participants may not even notice that their eyes are being monitored. In fact, eye-tracking can even be used with dogs when trained to keep their head in a chin rest (Somppi, Törnqvist, Hänninen, Krause, & Vainio, 2012). Participants don't have to perform any meta-linguistic tasks such as deciding which sounds they heard. Rather, all they have to do is monitor a visual scene and listen to the accompanying audio signal. They may or may not be asked to explicitly manipulate presented objects (e.g., on a computer screen by clicking on them or dragging them to different locations). An advantage of asking participants to perform a task is that response accuracy can be used to check whether participants engaged with the experiment. Eye-tracking hence lends itself to studying processing of a continuous speech signal.

One issue that remains is the extent to which the visual display that is presented to the listener may affect processing of the audio signal, that is, whether seeing the picture of a lion alters the word-recognition process due to the expectation to hear "lion". For a more detailed discussion of this issue and the use of visual world eye-tracking to study language processing beyond phonetic research questions we refer the reader to Huettig, Rommers, and Meyer (2011).

18.4. Recent Research

18.4.1 Segmental Processing: Cue Uptake, Context Effects, and Reduced Speech

Given that eye-tracking provides a continuous measure of lexical hypotheses—that is, which word the listener considers likely to be intended by the speaker at any given point in time—, it has been used to investigate how quickly different cues and contexts can influence spoken-word recognition. McMurray, Clayards, Tanenhaus, and Aslin (2008) were the first to use eye-tracking in that way and reported that cues that arrive earlier in the audio signal are also used earlier. For instance, they showed that for word-initial stop voicing distinctions in English (e.g., *peach* – *beach*) voice onset time (VOT) in word-initial stops is used earlier than the duration of the following vowel (which also cues stop voicing distinctions). This was reflected in earlier looks to the picture of a peach when the VOT was consistent with a voiceless /p/ (i.e., long VOT) than when the vowel duration was consistent with a voiceless /p/ (i.e., short vowel duration).

Reinisch and Sjeps (2013) tested whether acoustic cues that become available on the same segment (here duration and formant values cueing the Dutch contrast /a:/-/ɑ/) would show differences in their use over time. They found that spectral cues to a vowel contrast tended to be used slightly earlier than duration cues, even though cue trading experiments suggested equal importance of both cues in an offline task. Eye-tracking hence provides a measure to track the uptake of different phonetic cues over time. Consequently, it has also been used to investigate how early the perceptual consequences of anticipatory coarticulation can be used to infer the identity of an upcoming segment. For example, Beddor et al. (2013) showed that American English listeners could use anticipatory nasalization in a vowel to predict an upcoming nasal sound (e.g., in *bend* vs. *bed*). This prediction was evident in earlier fixations on the referent containing the nasal when more of the vowel was nasalised, that is, the earlier nasalization started. Coarticulatory cues were also important in establishing how fast eye-tracking can follow the auditory signal, that is, how long it takes for linguistic information to influence eye movements. The prevalent assumption is that this takes 150-200 ms (e.g., Allopenna et al., 1998). Altmann (2011) claimed that eye fixations can reflect linguistic processing within only 100 ms, but this lower estimate is problematic because the audio signal involved naturally produced sentences but the analysis did not control for coarticulatory cues. A subsequent study (Salverda, Kleinschmidt, & Tanenhaus, 2014) in which listeners heard typical instructions such as "Click on the

…" found that coarticulatory cues in the schwa in *the* can lead to early target fixations. Thus Salverda et al.'s results confirmed the earlier 150-200 ms estimate.

Building on this line of research, many studies have used eye-tracking to investigate the relative time course for the use of context information in the interpretation of the running speech signal. These studies have generally found that context is used immediately, be it speech rate (Reinisch & Sjerps, 2013; Toscano & McMurray, 2015), lexical context (i.e., the "Ganong" effect, see Kingston, Levy, Rysling, & Staub, 2016), or perceptual learning about a given speaker's idiosyncrasies (Mitterer & Reinisch, 2013).

Eye-tracking also has been used to estimate how listeners can revise their lexical hypotheses when there is a mismatch between initial acoustic cues and information about the identity of a word that becomes subsequently available. McMurray et al. (2009) manipulated onset consonants in long words such as *parakeet* such that they were closer to /b/ than to the canonical /p/ and hence initially matched competitors such as *barricade*. Results showed that the more /b/-like the initial sound was, the slower listeners were to revise their initial hypotheses (i.e., *barricade*) - with a close and gradient link between the acoustics of the initial sound and listeners' fixation patterns. This demonstrates that listeners modulate lexical competition based on fine phonetic detail, and contrasts with classical claims about categorical perception (e.g., Liberman, Harris, Hoffman, & Griffith, 1957). More recently, Brown-Schmidt and Toscano (2017) showed that listeners remain uncertain even beyond the word level. Using a continuum between the pronouns *he* and *she*, they found that even several words later, participants' fixations to male vs. female referents were guided by the acoustics of the word initial pronoun in a gradient fashion. This also contrasts with the winner-takes-all strategy of many computational models of perception. In a similar vein, Dahan and Tanenhaus (2004) investigated how listeners integrate sentential context and phonetic detail in Dutch spoken-word recognition. They used visual displays with pictures of two cohort competitors (such as Dutch *bot* "bone" and *bok* "goat") in which sub-phonemic coarticulatory cues to the last consonant either matched or mismatched through cross-splicing. For example, the onsets /bɔ/ of *bot* vs. *bok* were exchanged such that the coarticulatory cues in the vowels no longer matched the cues for stop release (i.e., /t/ vs. /k/). Although the stop release of the final consonant determined target recognition, mismatched coarticulatory cues delayed fixations on the target. In addition, the words were presented in either neutral or semantically biasing sentence frames[1]. Results showed that sentence context strongly

[1] Example sentences would be *Nog nooit klom een bok zo hoog* "Never before climbed a goat so high" vs. *Nog nooit is een bok zo hoog* geklommen "Never before has a goat climbed so high". The main verb is underlined, the target is *bok* in both cases.

influenced target activation as there were fewer looks to *bot* "bone" if the sentence favoured *bok* "goat". However, despite a mismatching semantic context, listeners did again consider the competitor *bot* if coarticulatory cues temporarily favoured the competitor (i.e., in the cross-spliced condition). Top-down knowledge hence does not seem to override bottom-up input.

Because eye-tracking can reveal the consideration of lexical candidates as the speech signal unfolds, it has also been used to investigate how listeners deal with casual speech processes. Several studies have investigated the use of context in compensation for phonological assimilations (Clayards, Niebuhr, & Gaskell, 2015; Gow & McMurray, 2007; Mitterer, Kim, & Cho, 2013) and showed how both, potentially incomplete neutralization and phonological context co-determine the recognition of the intended word. Given instructions such as "Click on the gree$^{n}_{m}$ …" where the nasal in *green* is the result of an incomplete coronal place assimilation, listeners use this information to look at potential referents that may trigger coronal place assimilation (i.e., *boat* but not *dog*). Effects of phonological context are reflected when participants hear phrases such as *ca$^{t}_{p}$ box* versus *ca$^{t}_{p}$ drawing* and see pictures of a cat and a cap in the visual display. More looks to the picture of the cap are found if the following word can trigger coronal place assimilation (as *box* but not *drawing* can).

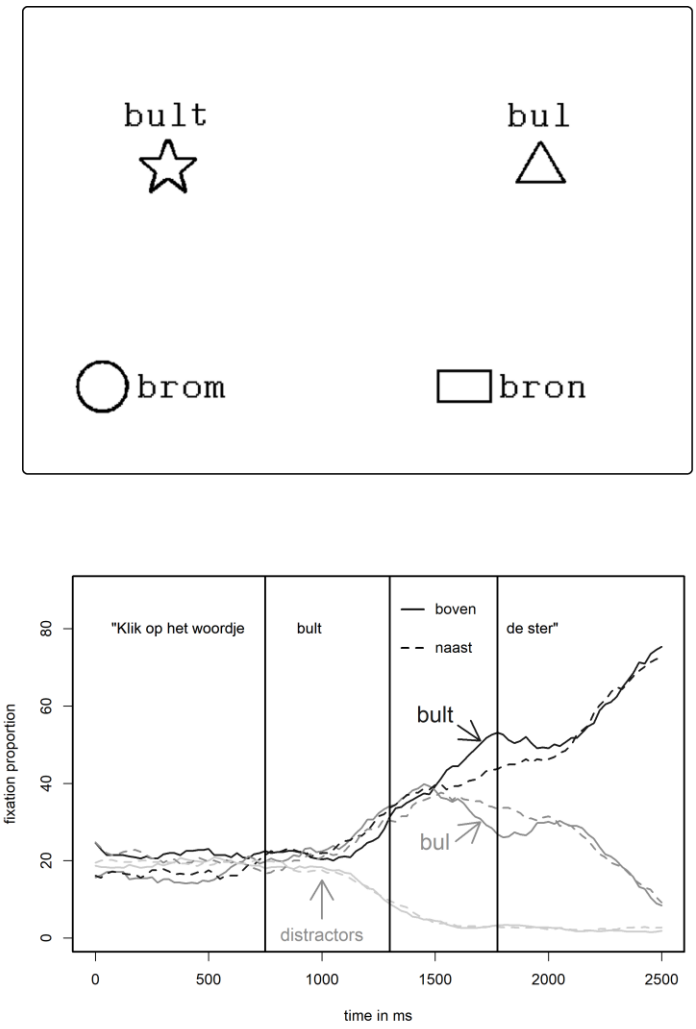

Figure 18.1: Example of an experimental display used in Mitterer and McQueen (2009) and the results obtained. When hearing the lead-in part of the sentence (starting at 0 ms and ending at the first vertical line), all words have roughly the same likelihood of being fixated on. When the target is heard, participants focus on the target or the competitor and away from the distractors. Finally, when the object name indicates the target (i.e., *bult* "hump") they look at it and away from the competitor *bul* ("diploma"). When the preposition *boven* ("above") is heard they look more towards the /t/-final word than when the preposition *naast* ("next to") is heard, reflecting the probability of /t/ deletion in these cases.

[INSERT FIGURE 18.1 ABOUT HERE]

Eye-tracking has also been used to investigate stronger forms of reductions, in which one or more segments in a word are not present at all. Mitterer and McQueen (2009) used displays such as in Figure 18.1 to investigate how listeners deal with /t/-deletion. The targets to be clicked on were written words—some of which differed only in the presence or absence of a final /t/— that were

accompanied by shapes (e.g., a triangle or a star). The instruction were Dutch sentences such as *Klik op het woordje bul/bult boven/naast de ster* ("Click on the word diploma/hump above/next to the star"). Figure 18.1 shows a display from this study as well as the results for /t/-final targets. The critical aspect is that the displays contained both (written) words *bul* and *bult* but only one of them (*bult*) was above the shape that was mentioned in the sentence (here: a star) while the other word (*bul*) was above a different shape (triangle). Since the /t/ in *bult* can be deleted, the instruction was ambiguous up until participants got the information about the shape. However, at this point the target became unambiguous since only the word *bult* was above a star so that *bul* was no longer the possible target. This design has two advantages: first, participants do not need to focus on phonetic detail to make their choice about the target, because the sentence context indicates the target unambiguously. Second, it tests to what extent phonetic detail and phonetic context matter before the sentence is semantically disambiguated. The results showed that listeners fixated on the /t/-final word prior to disambiguation more frequently if the signal contained phonetic detail consistent with a deleted /t/ and also if the following context was a labial (/b/ in *boven*, Engl., "above"). Since this pattern reflects previously reported production patterns with more frequent deletion in labial contexts in Dutch (Mitterer & Ernestus, 2006), this suggested that listeners use their implicit knowledge about reduction patterns in their native language to recognise words. Interestingly, this effect was not apparent in a phonetic identification task (Mitterer & Ernestus, 2006) which suggests that diverting attention away from the acoustic signal may in fact strengthen listeners' implicit reliance on it.

Eye-tracking has also been used to estimate how reductions slow word recognition. In such experiments, participants are instructed to click on one of four words presented on the screen if one of them occurs in the sentence that they hear. This allows the presentation of a wide variety of sentences and even the use of stimuli from spontaneous-speech corpora. Results from these studies showed deletion costs, that is, fewer early fixations on the target word when it was reduced than when it was produced in its full form. These deletion costs were smaller for high-frequency words and syntactically predictable words (Viebahn, Ernestus, & McQueen, 2015) but independent of the extent of discourse context (Brouwer, Mitterer, & Huettig, 2013). Mitterer and Reinisch (2015) used eye-tracking to investigate glottal marking of vowel-initial words in German versus glottal-stop initial words in Maltese. According to most accounts of German phonology, the glottal stop is not part of the lexical representation of vowel-initial words. In Maltese, on the other hand, the glottal stop is a phoneme and is therefore part of the lexical representation of words. Deletion costs were measured by how quickly participants fixated on the picture of a target word (e.g., the German word *Elch* /ʔelç/ ("moose") or the Maltese word *qamar* /ʔɑmɑr/ ("moon") depending on whether the target word was produced with or without an initial glottal stop. The deletion costs were comparable in both

languages, indicating that the glottal stop is part of the lexical representation of German words that are typically referred to as "vowel-initial".

Eye-tracking has not only been used to investigate spoken-word recognition across languages but also across native speakers and learners of the same language. Several studies found asymmetric competition patterns between minimal pairs containing sound contrasts that learners tend to confuse (Cutler, Weber, & Otake, 2006; Weber & Cutler, 2004). In these studies, visual displays usually contain pictures of words that - due to the lack of perceptual differentiation - are potential cohort competitors for second language (L2) learners. For instance, the English words *rocket* and *locker* are cohort competitors for Japanese learners of English when they fail to distinguish between the sounds /l/ and /ɹ/. Results showed that upon hearing *locker,* listeners indeed looked at both pictures – a locker and a rocket - even before the word offset would have allowed disambiguation. However, upon hearing *rocket*, there were fewer looks to the competitor *locker* than vice versa. That is, unlike native listeners' symmetric fixations on a picture of a beaker and a beetle with the input [bi] as described earlier, for Japanese learners, words with /l/ and /ɹ/ - though confusable - were not equally good cohort competitors for each other. This can best be explained by assuming that Japanese listeners have a different lexical representation for the onset sound in the words *rocket* and *locker*, even though they seem to have trouble distinguishing [ɹ] and [l] at the phonetic level (Cutler et al., 2006). This interpretation of course raises the question of how, given difficulties in distinguishing such sounds, listeners actually manage to learn which segment is part of which word. Eye-tracking studies suggest that learning may be based on orthographic and visual-articulatory cues (Escudero, Hayes-Harb, & Mitterer, 2008; Llompart & Reinisch, 2017). These studies used the English vowel contrast /æ/-/ɛ/ which is difficult for Dutch and German learners (see, e.g., Weber & Cutler, 2004). Participants learned to associate nonce words containing this contrast (e.g., [tændɛk] vs. [tɛnsəɹ]) with made-up objects. When learning object-word associations was based on audio only there was symmetric competition, which suggests that the critical vowels could not be differentiated. In other conditions, listeners were provided with additional orthographic or visual-articulatory cues when learning the association between the nonsense shapes and the nonwords. This allowed them to learn which vowel was used in which word, and as a result competition was asymmetric. This suggests that asymmetric competition patterns in an L2 can be used as a diagnostic to assess whether learners are able to differentiate difficult L2 sound contrasts at the lexical level. Importantly, eye-tracking is quite powerful because it can show how strongly words are activated when listening to an L2, and the results do not always follow the patterns that would be expected based on data from identification tasks.

A study by Hanulíková and Weber (2012) further showed that L2 listeners are influenced by their own production choices: While both Dutch and German learners of English struggle to produce

a dental fricative /θ/, the most frequent substitution for Dutch speakers is [t] but for German speakers it is [s]. When confronted with [s], [f], and [t] as substitutions for /θ/-initial words (i.e., *theft* produced as *seft, feft,* and *teft*), each listener group was found to look most at the /θ/-initial words when the stimulus contained the preferred substitution from their own L2 variety.

4.2 Suprasegmental Processing

Eye-tracking has also been used successfully to assess the role and timing of the uptake of different types of prosodic information. Studies on tone languages suggest that lexical access in native speakers of Mandarin is strongly and immediately influenced by both, segmental and tonal information (e.g., Malins & Joanisse, 2010; Shen, Deutsch, & Rayner, 2013). Malins and Joanisse (2010), for example, found equal amounts of competition to a target word if the competitor was a full segmental match to the target but carried a different tone (*chuang1* "window" for the target *chuang2* "bed"; where 1 and 2 refer to the high level and rising tone, respectively) or if its segmental overlap was only word-initial but matched in tone (*qian2* "money" for the target *qiu2* "ball").

Word-level prosody in the form of suprasegmental cues to lexical stress have been studied in Dutch (Reinisch, Jesse, & McQueen, 2010; Reinisch & Weber, 2012), Italian (Sulpizio & McQueen, 2012), and English (Jesse, Poellmann, & Kong, 2017; Quam & Swingley, 2014). For example, Reinisch et al. (2010), tracked Dutch listeners' fixations on printed-word referent pairs of the type *OCtopus - ocTOber* (capitals are used to indicate stress placement here but were not presented to participants) while being instructed to click, for instance, on *octopus*. Results showed that listeners fixated on the target word (*octopus*) more than the segmentally overlapping competitor (*October*; note that in Dutch unstressed syllables are typically not reduced) even before any disambiguating segmental information could have been processed. That is, listeners were able to use suprasegmental lexical stress cues to disambiguate target and competitor before they became segmentally distinct.

Importantly, eye-tracking experiments like this have shown that listeners of different native languages differ in how they weight suprasegmental cues to stress and in the extent to which they are influenced by language-specific default patterns. For instance, Italian listeners interpret segmentally ambiguous words by default as carrying penultimate stress unless acoustic cues, specifically amplitude and F0, support an antepenultimate stress pattern (Sulpizio & McQueen, 2012). Reinisch and Weber (2012) showed that stress perception remains plastic, as Dutch listeners re-weighted the importance of stress information when a non-native speaker produced stress errors consistently. Brown, Salverda, Dilley, and Tanenhaus (2015) demonstrated effects of the rhythmic pattern of context sentences preceding the target word in English with more looks to targets with an iambic stress pattern when the context sentence had an iambic pattern.

Another type of word-level prosody that has been shown to modulate lexical access using an eye-tracking paradigm is prosodic cues to word boundaries. Salverda, Dahan, and McQueen (2003) showed that listeners use sub-phonemic lengthening to predict upcoming word boundaries: the longer the duration of a syllable like *ham*, the more likely listeners were to interpret the sequence as the monosyllabic word *ham* relative to the di-syllabic competitor *hamster*. This result could be explained either by assuming an episodic account in which listeners store acoustic details of how given words are produced or, alternatively, by assuming abstract storage and a concurrent prosodic analysis that influences lexical activation. To distinguish between these two explanations, Shatzman and McQueen (2006a) taught participants to associate new mono- and disyllabic nonce words that overlapped in their first syllable with novel shapes. Crucially, the duration of the first syllable in the mono- and disyllabic words was identical during exposure, which means that an episodic account would predict no influence of syllable duration on eye-fixations. However, as for real words, effects of syllable duration were found for these newly learned words. This indicates that it is not episodic storage of acoustic detail but rather a prosodic analysis that can explain the effects found by Salverda et al. (2003).

Listeners also make immediate, online use of segment durations at word boundaries to segment the speech stream into words. Shatzman and McQueen (2006c, 2006b; Reinisch, Jesse, & McQueen, 2011) showed that a segment /s/ at a word boundary in sequences such as Dutch *eens peer* vs. *een speer* ("once pear" - "one spear") is interpreted as belonging to the preceding word if it is short but as belonging to the onset of the second word if it is long. That is, a stop-initial target (*peer*) was fixated on faster following a short than a long /s/.

Eye-tracking studies have also shown that the duration of word-boundary segments is evaluated relative to the speaking rate of the context (Reinisch et al., 2011) and that the interpretation of segmental lengthening depends on the position of the target word in the prosodic phrase (Brown, Salverda, Dilley, & Tanenhaus, 2011; Salverda et al., 2007). For example, in a sentence such as "Put the *cap* next to the square" where the monosyllablic target *cap* occurs in sentence-medial position, there are more fixations on a bi-syllabic competitor *captain* than on another monosyllabic word *cat* (Salverda et al., 2007). However, the reverse was found in sentence-final position ("Now click on the *cap*") where the monosyllabic competitor *cat* is the stronger competitor. That is, in sentence-final position the prosodic word boundary combines with the prosodic phrase boundary that favours the prosodically matching competitor (*cap*).

Another important issue in the processing of prosody that has been addressed with eye-tracking is the use of sentence intonation to predict or disambiguate upcoming referents in an utterance. The typical paradigm here consists of pairs of instructions in which the first introduces

certain referents or referent sets. The second instruction is the critical one, in which the timecourse of referent resolution is measured depending on the information status of the target (e.g., given, new) and the prosody (e.g., accented, unaccented) of the sentence.

For example, Dahan, et al. (2002) asked listeners to move objects on a screen given instructions such as "Put the *candle* above the square; now put the *candle/candy* below the circle" where fixations on the pictures of the candy and the candle (the potential referents) were monitored during the second part of the instruction. When the first syllable of the target contained a pitch accent, listeners tended to anticipate the new object (*candy*) as the intended referent. However, when the target was unaccented they anticipated another mention of the given object (*candle*) as reflected in more looks to the respective objects (see also Arnold, 2008; Brown, Salverda, Gunlogson, & Tanenhaus, 2015; Dahan et al., 2002).

Similar effects for given vs. new information have been found when the referent set consisted of coloured objects that formed contrast pairs. For example, Ito and Speer (2008, 2011) asked listeners to decorate holiday trees with different objects in a variety of colours. They demonstrated that an accent on the adjective in the first instruction (e.g., "hang the BLUE *bulb*" - with a contrastive pitch accent on *blue*) leads to the expectation of a contrastive colour of the same object in the upcoming sentence, as evidenced by increased fixations on bulbs in other colours relative to fixations on other objects (see also Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Weber, Braun, & Crocker, 2006).

In addition to the role of prosody in disambiguating the referent of a sentence, studies have also examined its role in resolving temporary syntactic ambiguities (e.g., Nakamura, Arai, & Mazuka, 2012, for Japanese ; Snedeker & Trueswell, 2003, for English ; Weber, Grice, & Crocker, 2006, for German). In Weber et al., for instance, German listeners were exposed to sentences in which the first noun phrase could either be the sentence's subject or object. If the intonation on the first noun phrase favoured a subject interpretation, listeners anticipated the second noun phrase to refer to a patient (i.e., object) which was reflected in their direction of fixations on possible patients within the visual scene before the respective word had been heard. Moreover, eye-tracking has also been used to test whether listeners make use of different types of pitch accent rather than simply responding to the presence vs. absence of an accent (e.g., Chen, den Os, & de Ruiter, 2007; Heeren, Bibyk, Gunlogson, & Tanenhaus, 2015; Watson, Tanenhaus, & Gunlogson, 2008). Watson et al. (2008), for example, showed that a word with a high pitch accent (H* in the ToBI annotation, Beckman & Hirschberg, 1994) was interpreted as referring to less salient information independently of whether the referent was given in the discourse context. L+H*, in contrast, appeared to have a narrower interpretation in signalling contrast only.

18.4.3 Gaze in Interaction

Eye-tracking has been used to investigate where and when participants look at a speaker and how this might influence speech perception. Gaze was traditionally a topic in conversation analysis (e.g., Rossano, Brown, & Levinson, 2009) and is concerned with how interlocutors look at each other during the flow of conversation. This kind of work often does not make use of a dedicated eye-tracker but simply analyses gaze direction in video recordings. An interesting new avenue for this kind of work is the recent introduction of eye-tracking glasses, which may allow a more accurate tracking in future studies of eye movements in conversation (see *Best Practice* section).

Another issue regarding gaze in interaction is how it affects the use of visual versus auditory cues for speech perception. Here, the expectation might be that fixations on the mouth area lead to a stronger visual influence. However, audiovisual speech perception seems to be quite independent of the actual gaze position as long as the face is within twenty degrees of the fixation position (Hisanaga, Sekiyama, Igasaki, & Murayama, 2016; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Given that visual attention and gaze position do not need to coincide (see, e.g., van der Heijden, 1992), these findings only show that audiovisual integration is independent of gaze but not that audiovisual integration is independent of visual attention. In fact, there is evidence that attention influences audiovisual integration; the effectiveness of visual cues depends on how much visual attention there is for the face (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Mitterer & Reinisch, 2017).

While gaze position close to the mouth of the speaker does not seem to be important for the use of visual speech cues once they have been learned, the learning itself may depend on gaze position. Hisanaga et al. (2016) found that Japanese listeners are less influenced by visual speech signals than English listeners and also fixate less on the mouth than English listeners. Note, however, that earlier research had established that the exact fixation position is not crucial for a visual influence to arise (Vatikiotis-Bateson et al., 1998). The most likely account for this pattern is that Japanese listeners do not learn strong associations between visual cues and phonetic categories because they habitually do not attend to the mouth area during interactions. However, for listeners who have learned this association, the actual gaze position does not matter. That is, gaze position has little influence on the use of visual speech signal in a given situation.

18.5. Best Practice for Teaching and Learning

In this section, we will discuss which issues need to be considered and/or taught when implementing a visual-world eye-tracking study. Given the specificities of eye-tracking designs, teaching the method is suitable not only to discuss what types of research questions on real-time speech comprehension can be addressed with this method, but it can also be used to teach various aspects of experimental design, computer-programming and/or statistical analyses in a problem-

based manner. Most of the studies discussed above made use of screen-based eye-tracking. In this situation, the eye-tracker returns screen coordinates which then can be related to the visual display, that is, to categorise the fixation position as a fixation on the target, competitor, or some other referent. Alternatively, the distance between the centre of gravity of the object and the fixation position can be used as the dependent variable (Mitterer & McQueen, 2009; Nixon, van Rij, Mok, Baayen, & Chen, 2016). Converting the raw output of screen coordinates to fixations or distances provides a good opportunity to teach computer-programming.

When planning a screen-based visual-world study, one of the first decisions is whether the visual referents should be pictures or written words. If possible, pictures provide a more natural choice and allow for the comparison of diverse populations, including those who may have trouble reading. However, using pictures can also be problematic. For instance, in many languages it would be extremely difficult to find a sufficient number of minimal word pairs that differ in the place of articulation of a word-final nasal, with both members being picturable. It may further be interesting to discuss with students what other possible drawbacks picture referents may have, for instance, semantic relatedness (Huettig & McQueen, 2007) or similarities in colour or shape (Huettig & Altmann, 2007). When only a small number of suitable items can be found, it is important to consider that statistical power will remain low, even with a very large number of participants (cf. Westfall, Kenny, & Judd, 2014). In such cases, using written words may be the better choice. While using written words opens the door to possible orthographic effects, such effects appear to play less of a role when participants have ample time to preview the display (Salverda & Tanenhaus, 2010). Whether using pictures or words, lexical properties such as word frequency, neighbourhood density, etc. have been shown to influence the pattern of activation (Magnuson, Dixon, Tanenhaus, & Aslin, 2007) and these factors should be taken into account wherever possible, be it in stimulus selection or in the statistical analysis.

Another practical issue that is not to be underestimated is the complexity of counterbalancing all relevant factors in a visual-world study. For experiments using identification or discrimination tasks, a trial has only a few variables (one or two sound files and a potentially variable inter-stimulus interval). A trial for a visual-world study, in contrast, involves at least ten variables: An auditory stimulus, information about the onset of the critical word in that auditory stimulus, four visual stimuli and their respective position on the screen. The position of referents on the screen is important because participants with experience with a left-to-right writing system have the tendency to scan the visual field from the top-left to the bottom-right after display onset (see Nixon et al., 2016). An (accidental) overrepresentation of targets in the top left corner could hence lead to an unusually quick convergence of fixations on the target. In our experience, experiments have had to be repeated since

an apparently early convergence on the target in one condition was in fact caused by an overrepresentation of top-left target positions for that condition. This can be avoided by counterbalancing target and competitor positions within each condition or at least over the whole experiment.[2] Another remedy is to have ample display preview before the auditory stimulus starts, as this has also been shown to alleviate the top-left bias to some extent (Nixon et al., 2016).

It is also important to note that although eye-tracking provides a continuous measure of word activation over time, any fixation by a participant on a given trial is a discrete event. The impression of continuity arises only by averaging over many trials and participants. Moreover, although many questions addressed with eye-tracking are about the timing of the use of different types of phonetic information, assessing a specific point in time at which something happens is not a trivial issue. In many of the examples discussed above, eye-tracking was used to show that information X was used before information Y. However, statements about the temporal order of effects are not trivial to support statistically. Some of the methods for analysing the timecourse of fixations (Barr, 2008; Mirman, Dixon, & Magnuson, 2008) in fact require an a-priori decision about when effects are likely to start. Estimating the onset of an effect in an eye-tracking experiment has often borrowed techniques from the event-related-potential field. These include jack-knife methods, in which the onset of an effect is estimated based on aggregated data minus one participant (McMurray et al., 2008; Mitterer & Reinisch, 2013).

In general, the analysis of eye-tracking data is not straightforward because the data are categorical and often require generalization over participants and items. The field is too much in flux at the time of writing to provide positive advice—new data analysis methods are proposed nearly weekly in the current psycholinguistic literature—but it is possible to give some words of caution. It is highly problematic to analyse raw fixation proportions with an analysis of variance (or a linear mixed-effect model for that matter). At the very least, a logit or probit transformation of the raw data should be used; this means that the data have similar properties to the d-prime measure from signal-detection theory, which is generally thought to be unproblematic for analysis with parametric methods. However, such transformations are not always ideal (Donnelly & Verkuilen, 2017) and an alternative would be to analyse whether there is a fixation or not in a given time window. In this way,

[2] In our own labs, custom randomization routines are usually written (in Perl, Python, or R, for example) to randomise both trial order and display positions for each participant separately to minimise any such bias. We also can recommend analysing whether the output of such a routine does indeed have the desired properties (e.g., an equal or roughly equal number of top-left target position within each experimental condition over all trials). Writing such short programmes is well suitable for students to practice basic programming skills.

the data can be analysed with a generalised linear-mixed effect model with a logit-link function (see, e.g., Brown-Schmidt & Toscano, 2017). Note also that "Object" (target versus competitor) should not be used as an independent variable. Most analyses assume some form of independence of the outcomes for different levels of the independent variable. However, having object as an independent variable violates this assumption because a fixation on the target means that there is no fixation on the competitor. For such cases, it is advisable to generate a *dependent variable* that reflects the relative attractiveness of the two objects, such as the logged ratio (i.e., log([fixations to target]/[fixations to competitor]) with some correction for cases in which this ratio is zero or undefined.

Another potential issue is the determination of time windows to be analysed. It will often make sense to let this time window start 200 ms after the onset of the critical word, but finding an appropriate end of the window, or the timing of separate windows can be problematic and increases the researchers' degree of freedom. Moreover, the outcomes for successive time windows are hardly ever completely independent, because fixations carry over from one window into the next. In some cases, it may be possible to add filler trials that will help to determine when participants process which part of a sentence, and this information may be used to determine the time window for the experimental trials (see, e.g., Mitterer & McQueen, 2009).

18.6. Future Directions

Given that basically all studies discussed in this chapter made use of stationary lab set-ups we expect that in the future the use of free-viewing eye-trackers, for example in the form of glasses, will increase. However, currently, the use of glasses comes at the cost of tedious post-processing routines because such trackers simply deliver a camera view that roughly corresponds to the participant's view plus a fixation position in that view. Unless some automatic image parsing is possible, this means that each fixation must be coded manually. Moreover, most free-viewing eye-trackers work with relatively low sampling frequencies (around 30Hz). Given the speed of eye-movements, such a sampling rate is too slow to answer questions about timing.

Summing up what we learned from the eye-tracking method so far, it provides an online measure of how listeners interpret the unfolding speech signal. It has proven useful to address many issues on listeners' use and temporal uptake of phonetic information during spoken word recognition. Although we have highlighted the complexity of setting up eye-tracking experiments, the large number of variables for a given trial also is a strength of the paradigm. There are few bounds on the creativity of the researcher in combining visual displays and speech signals. In the last two decades, eye-tracking research has shown that listeners are sensitive to fine phonetic detail and make use of sub-segmental, segmental, and prosodic information as soon as it becomes available. Moreover,

phonetic context is used immediately to interpret and sometimes even anticipate upcoming information and retained in some form if the context requires a reinterpretation. Understanding how listeners modulate lexical competition in real-time has to a large extent been driven by the appearance and refinement of the eye-tracking method.

18.7. 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*(4), 419–439.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual Integration of Speech Falters under High Attention Demands. *Current Biology*, *15*(9), 839–843. https://doi.org/10.1016/j.cub.2005.03.046

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

Arnold, J. E. (2008). THE BACON not the bacon: How children and adults understand accented and unaccented noun phrases. *Cognition*, *108*(1), 69–99. https://doi.org/10.1016/j.cognition.2008.01.001

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

Beckman, M., & Hirschberg, J. (1994). *The ToBI annotation conventions*. Colombus, OH: Ohio State University.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America*, *133*(4), 2350–2366. https://doi.org/10.1121/1.4794366

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Can hearing puter activate pupil? Phonological competition and the processing of reduced spoken words in spontaneous conversations. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2193–2220. https://doi.org/10.1080/17470218.2012.693109

Brouwer, S., Mitterer, H., & Huettig, F. (2013). Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics*, *34*, 519–539. https://doi.org/10.1017/s0142716411000853

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin & Review*, *18*(6), 1189–1196. https://doi.org/10.3758/s13423-011-0167-9

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 306–323. https://doi.org/10.1037/a0038689

Brown, M., Salverda, A. P., Gunlogson, C., & Tanenhaus, M. K. (2015). Interpreting prosodic cues in discourse context. *Language, Cognition and Neuroscience*, *30*(1–2), 149–166. https://doi.org/10.1080/01690965.2013.862285

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, *0*(0), 1–18. https://doi.org/10.1080/23273798.2017.1325508

Chen, A., den Os, E., & de Ruiter, J. P. (2007). Pitch accent type matters for online processing of information status: Evidence from natural and synthetic speech. *The Linguistic Review*, *24*(2–3), 317–344. https://doi.org/10.1515/TLR.2007.012

Clayards, M., Niebuhr, O., & Gaskell, M. G. (2015). The time course of auditory and language-specific mechanisms in compensation for sibilant assimilation. *Attention, Perception, & Psychophysics*, *77*(1), 311–328. https://doi.org/10.3758/s13414-014-0750-z

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X

Creel, S. C., Tanenhaus, M. K., & Aslin, R. N. (2006). Consequences of lexical stress on learning an artificial lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 15–32. https://doi.org/10.1037/0278-7393.32.1.15

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*, 269–284. https://doi.org/10.1016/j.wocn.2005.06.002

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 498–513.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292–314. https://doi.org/10.1016/S0749-596X(02)00001-3

Donnelly, S., & Verkuilen, J. (2017). Empirical logit analysis is not logistic regression. *Journal of Memory and Language*, *94*, 28–42. https://doi.org/10.1016/j.jml.2016.10.005

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*, 345–360. https://doi.org/10.1016/j.wocn.2007.11.002

Gow, D. W., & McMurray, B. (2007). Word recognition and phonology: The case of English coronal place assimilation. In J. Cole & J. Hualde (Eds.), *Laboratory Phonology 9* (pp. 173–200). New York: Mouton de Gruyter.

Hanulíková, A., & Weber, A. (2012). Sink positive: Linguistic experience with th substitutions influences nonnative word recognition. *Attention, Perception, & Psychophysics*, *74*(3), 613–629. https://doi.org/10.3758/s13414-011-0259-7

Heeren, W. F. L., Bibyk, S. A., Gunlogson, C., & Tanenhaus, M. K. (2015). Asking or Telling – Real-time Processing of Prosodically Distinguished Questions and Statements. *Language and Speech*, *58*(4), 474–501. https://doi.org/10.1177/0023830914564452

Hisanaga, S., Sekiyama, K., Igasaki, T., & Murayama, N. (2016). Language/Culture Modulates Brain and Gaze Processes in Audiovisual Speech Perception. *Scientific Reports*, *6*, srep35265. https://doi.org/10.1038/srep35265

Huettig, F., & Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, *15*(8), 985–1018. https://doi.org/10.1080/13506280601130875

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460–482. https://doi.org/10.1016/j.jml.2007.02.001

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

Isaacs, A. M., & Watson, D. G. (2010). Accent detection is a slippery slope: Direction and rate of F0 change drives listeners' comprehension. *Language and Cognitive Processes*, *25*(7–9), 1178–1200. https://doi.org/10.1080/01690961003783699

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*(2), 541–573. https://doi.org/10.1016/j.jml.2007.06.013

Ito, K., & Speer, S. R. (2011). Semantically-Independent but Contextually-Dependent Interpretation of Contrastive Accent. In S. Frota, G. Elordieta, & P. Prieto (Eds.), *Prosodic Categories: Production, Perception and Comprehension* (pp. 69–92). Springer Netherlands. https://doi.org/10.1007/978-94-007-0137-3_4

Jesse, A., Poellmann, K., & Kong, Y.-Y. (2017). English Listeners Use Suprasegmental Cues to Lexical Stress Early During Spoken-Word Recognition. *Journal of Speech, Language, and Hearing Research*, *60*(1), 190–198. https://doi.org/10.1044/2016_JSLHR-H-15-0340

Kingston, J., Levy, J., Rysling, A., & Staub, A. (2016). Eye movement evidence for an immediate Ganong effect. *Journal of Experimental Psychology. Human Perception and Performance*, *42*(12), 1969–1988. https://doi.org/10.1037/xhp0000269

Kong, Y.-Y., & Jesse, A. (2017). Low-frequency fine-structure cues allow for the online use of lexical stress during spoken-word recognition in spectrally degraded speech. *The Journal of the Acoustical Society of America*, *141*(1), 373–382. https://doi.org/10.1121/1.4972569

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

Llompart, M., & Reinisch, E. (2017). Articulatory information helps encode lexical contrasts in a second language. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(5), 1040–1056. https://doi.org/10.1037/xhp0000383

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The Dynamics of Lexical Competition During Spoken Word Recognition. *Cognitive Science*, *31*(1), 133–156. https://doi.org/10.1080/03640210709336987

Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language*, *62*(4), 407–420. https://doi.org/10.1016/j.jml.2010.02.004

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, *15*(6), 1064–1071.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91. https://doi.org/10.1016/j.jml.2008.07.002

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Mitterer, H., & Ernestus, M. (2006). Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch. *Journal of Phonetics*, *34*(1), 73–103. https://doi.org/10.1016/j.wocn.2005.03.003

Mitterer, H., Kim, S., & Cho, T. (2013). Compensation for complete assimilation in speech perception: The case of Korean labial-to-velar assimilation. *Journal of Memory and Language*. https://doi.org/10.1016/j.jml.2013.02.001

Mitterer, H., & McQueen, J. M. (2009). Processing reduced word-forms in speech perception using probabilistic knowledge about speech production. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 244–263. https://doi.org/10.1037/a0012730

Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language*, *69*(4), 527–545. https://doi.org/10.1016/j.jml.2013.07.002

Mitterer, H., & Reinisch, E. (2015). Letters don't matter: No effect of orthography on the perception of conversational speech. *Journal of Memory and Language*, *85*, 116–134. https://doi.org/10.1016/j.jml.2015.08.005

Mitterer, H., & Reinisch, E. (2017). Visual speech influences speech perception immediately but not automatically. *Attention, Perception, & Psychophysics*, *79*(2), 660–678. https://doi.org/10.3758/s13414-016-1249-6

Nakamura, C., Arai, M., & Mazuka, R. (2012). Immediate use of prosody and context in predicting a syntactic structure. *Cognition*, *125*(2), 317–323. https://doi.org/10.1016/j.cognition.2012.07.016

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125. https://doi.org/10.1016/j.jml.2016.03.005

Quam, C., & Swingley, D. (2014). Processing of lexical stress cues by young children. *Journal of Experimental Child Psychology*, *123*, 73–89. https://doi.org/10.1016/j.jecp.2014.01.010

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *The Quarterly Journal of Experimental Psychology*, *63*(4), 772–783.

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.

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.

Reinisch, E., & Weber, A. (2012). Adapting to suprasegmental lexical stress errors in foreign-accented speech. *The Journal of the Acoustical Society of America*, *132*(2), 1165–1176.

Rossano, F., Brown, P., & Levinson, S. C. (2009). Gaze, questioning and culture. In J. Sidnell (Ed.), *Conversation analysis: Comparative perspectives* (pp. 187–249). Cambridge University Press.

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. https://doi.org/10.1016/S0010-0277(03)00139-2

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically-modulated sub-phonetic variation on lexical competition. *Cognition*, *105*(2), 466–476. https://doi.org/10.1016/j.cognition.2006.10.008

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*(1), 145–163. https://doi.org/10.1016/j.jml.2013.11.002

Salverda, A. P., & Tanenhaus, M. K. (2010). Tracking the time course of orthographic information in spoken-word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(5), 1108–1117. https://doi.org/10.1037/a0019901

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. https://doi.org/10.1016/S0010-0277(99)00025-6

Shatzman, K. B., & McQueen, J. M. (2006a). Prosodic Knowledge Affects the Recognition of Newly Acquired Words. *Psychological Science*, *17*(5), 372–377. https://doi.org/10.1111/j.1467-9280.2006.01714.x

Shatzman, K. B., & McQueen, J. M. (2006b). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics*, *68*(1), 1–16. https://doi.org/10.3758/BF03193651

Shatzman, K. B., & McQueen, J. M. (2006c). The modulation of lexical competition by segment duration. *Psychonomic Bulletin & Review*, *13*(6), 966–971. https://doi.org/10.3758/BF03213910

Shen, J., Deutsch, D., & Rayner, K. (2013). On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements. *The Journal of the Acoustical Society of America*, *133*(5), 3016–3029. https://doi.org/10.1121/1.4795775

Shockey, L. (2003). *Sound patterns of spoken English*. Cambridge, MA: Blackwell.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130. https://doi.org/10.1016/S0749-596X(02)00519-3

Somppi, S., Törnqvist, H., Hänninen, L., Krause, C., & Vainio, O. (2012). Dogs do look at images: eye tracking in canine cognition research. *Animal Cognition*, *15*(2), 163–174. https://doi.org/10.1007/s10071-011-0442-1

Sulpizio, S., & McQueen, J. M. (2012). Italians use abstract knowledge about lexical stress during spoken-word recognition. *Journal of Memory and Language*, *66*(1), 177–193. https://doi.org/10.1016/j.jml.2011.08.001

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, *268*(5217), 1632–1634.

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 der Heijden, A. H. C. (1992). *Selective attention in vision*. New York: Routledge.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisualspeech perception. *Perception & Psychophysics*, *60*(6), 926–940. https://doi.org/10.3758/BF03211929

Viebahn, M. C., Ernestus, M., & McQueen, J. M. (2015). Syntactic predictability in the recognition of carefully and casually produced speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1684–1702. https://doi.org/10.1037/a0039326

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting Pitch Accents in Online Comprehension: H* vs. L+H*. *Cognitive Science*, *32*(7), 1232–1244. https://doi.org/10.1080/03640210802138755

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding Referents in Time: Eye-Tracking Evidence for the Role of Contrastive Accents. *Language and Speech*, *49*(3), 367–392. https://doi.org/10.1177/00238309060490030301

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*, 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Weber, A., Grice, M., & Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, *99*(2), B63–B72. https://doi.org/10.1016/j.cognition.2005.07.001

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology. General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014