The effect of speech degradation on the ability to track and predict turn structure in conversation

Céline Hidalgo, Isaïh Mohamed, Christelle Zielinski, Daniele Schön

Title: The effect of speech degradation on the ability to track and predict turn structure in conversation

Céline Hidalgo[1], Isaïh Mohamed[1], Christelle Zielinski[2], Daniele Schön[1]

[1] Aix Marseille Univ, Inserm, INS, Inst Neurosci Syst, Marseille, France

[2] Aix-Marseille Univ, Institute of Language, Communication and the Brain, France

Corresponding author: Céline Hidalgo, celine.hidalgo@univ-amu.fr

Professional Address: Institut de Neurosciences des Systèmes, Faculté de Médecine, bd Jean Moulin, 13005 Marseille, France

Telephone: 04 91 32 41 02

Abstract

Conversation represents a considerable amount of the daily language usage and plays an important role in language acquisition. In conversation, listeners simultaneously process their interlocutor's turn and prepare their own next turn. As such the turn-taking dynamics heavily relies on prediction. In other words, listeners avail prior knowledge to constrain both speech perception and production. Here we explored the relation between prediction and comprehension while watching two interlocutors having a conversation. We capitalize on gaze switch as a proxy of predictive behaviour to class dialogue turns as more or less well predicted and explore how this affects dialogue comprehension. Moreover, we study the extent to which speech degradation, by increasing the global uncertainty of the context, affects the relation between predictions, brain correlates of prediction errors (N400) and

global comprehension. Results show that 1) listeners direct gaze to the current speaker, in particular in challenging conditions, 2) gaze behaviour possibly affects the semantic processing of the upcoming turn (N400), 3) participants with a more efficient gaze predictive behaviour better solve semantic uncertainties at the turn onset, in particular in the most challenging listening condition. Our findings contribute to a better understanding of the relation between predictions, neural predictions errors and speech comprehension under challenging conditions.

Keywords: conversation, ERPs, eye-tracking, prediction, turn-taking.

Author contributions: Conceptualization C.H. and D.S.; Data curation C.H., C.Z. and I.M.; Formal Analysis C.H.; Funding acquisition D.S.; Project administration C.H.; Supervision D.S.; Visualization C.H, C.Z. and I.M; Writing – original draft C.H., I.M., C.Z. and D.S.

Author Note: In accordance with the Peer Reviewers' Openness Initiative (https://opennessinitiative.org, Morey, Chambers, Etchells, Harris, Hoekstra, Lakens, et al., 2016), all materials and scripts associated with this manuscript are available on https://osf.io/7dc6y/

Introduction

Speech perception is not the result of a real-time decoding of audio-visual information. It is a dynamic process of an anticipatory nature building on 1) the use of the linguistic and extra-linguistic context to generate hypotheses about the upcoming signal, i.e. we make predictions about what we are likely to hear and 2) a comparison of our predictions against what is actually perceived. Thus, when we listen to a conversation between two people, we anticipate, for example, the words that the speakers will produce according to the linguistic context that precedes their utterances. The more we know about this context, the fewer prediction errors we make and the faster and more accurate our understanding of their conversation. According to predictive coding theory (Clark, 2013), predictions would indeed be a default brain mode of operation for processing sensory information (Friston, 2005). To optimize speech perception, the brain would pre-activate representations of the expected speech input (Molinaro et al., 2016) at a phonological, lexical, semantic or even syntactic level (Pickering & Gambi, 2018). With increasingly accurate predictions, the brain would only need to compute the difference between sensory input and prediction, thereby decreasing the cost of perceptual processing and speeding up the understanding of the spoken message. A well-known fact in the comprehension literature is that the amount of contextual information, i.e. the increase in the precision of prediction, positively influences the speed of word recognition (Tyler & Wessels, 1983). This has been thoroughly studied using event-related potentials, by observing changes in the amplitude of a negative component. Indeed, words with low cloze probability engender a prediction error that is visible in an increase of the N400 amplitude. By contrast the N400 amplitude is reduced when perceiving speech in highly predictive contexts (see for a review Kutas & Federmeier 2011). However, until now, no study had shown a direct link between contextual predictions

and the integration of these predictions in the semantic comprehension process in a conversational context. This is due to the fact that, while prediction errors engender observable brain responses (e.g., N400 or Mismtach Negativity), this is not the case of predictions. Recently, Grisoni and colleagues (Grisoni et al., 2021) described an electrophysiological index of predictions called the "Semantic Prediction Potential", and showed, by measuring the N400 brain response during the perception of more or less contextually constrained sentences, the direct effect of predictions (SPP) on the processing of prediction errors (N400).

Another possibility to infer the ongoing predictions is to use behavioural measures. In the case of language, several studies have used for instance eye movements during reading (Rayner, 1978; Staub, 2015). Similarly, in speech perception, some authors relied on the implicit anticipatory gaze behaviour when listening to different people having a conversation (Casillas & Frank, 2017; Keitel et al., 2013). More precisely, listeners anticipate the end of the turn of the current speaker and switch gaze to the following speaker, before the actual onset of the new turn. This predictive behaviour is particularly important considering that the silence separating speech turns of conversing speakers only lasts a few hundred milliseconds (Levinson, 2016; Stivers et al., 2009). In considering turn-taking in conversation, predictions build on several linguistic cues such as prosody (Roberts et al., 2015), lexical-semantic cues (De Ruiter et al., 2006), syntactic structures (Selting, 1996) or pragmatic cues (Beňuš et al., 2011). When listening to a dialogue (as a third person), a similar behaviour possibly takes place via the generation of internal models, allowing a listener to anticipate in a precise temporal window the speakers' turn-taking.

The aim of the current study is to explore the links between predictions at the turn level, measured by gaze switch, and both local and global semantic processing when listening to a dialogue, as indexed by N400 to turn onset and comprehension score, respectively. Importantly we are interested in studying how these relations evolve when adding uncertainty to the context by degrading the input signal. Indeed, degrading the acoustic signal or adding noise or competing speakers, adds uncertainty and challenges speech comprehension (Mattys et al., 2012; Peelle, 2018). To this aim, we asked participants to watch videos of two people having a conversation manipulating different levels of speech degradation. We measured eye movements to estimate predictive gaze behaviour and used this measure to class turns as well predicted (early gaze switch) or poorly predicted (late gaze switch). We also measured EEG and used the N400 as a marker of prediction error at the turn onset and studied the extent to which this response is influenced by gaze behaviour. Finally, all participants responded to several open questions at the end of each dialogue, allowing estimating a global comprehension measure. Most importantly, we explored the relation between these different variables as a function of speech degradation and increasing uncertainty. We hypothesized that poorly predicted turns (late gaze switch) would engender a larger N400 response compared to well anticipated turns (early gaze switch). Importantly, this effect may change as a function of speech degradation and should be reduced at high speech degradation levels when uncertainty is highest. We also hypothesized that, if predictions are fully implicit and based on the understanding of the dialogue, then gaze switches should behave similarly to comprehension score and decrease from normal to moderately degraded and even more for highly degraded speech.

Materials and Methods

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. No part of the study procedures or analyses was preregistered prior to the research being conducted.

Participants

Forty-eight participants (27 females) were tested after being informed of the procedure of the study, which was approved by the Sud Méditerranée Ethics Committee (ID RCB: 2015-A01490-49). In absence of a known effect size, the sample size was chosen as rather large with respect to N400 studies (Šoškić et al., 2021). Inclusion and exclusion criteria: Participants were between 18 and 62 years old (mean = 32, SD = 6 years). They were native French speakers. The had normal or corrected to normal vision and normal hearing. years); they had no history of speech disorder or neurological disease based on self-report. Prior the experiment, the hearing thresholds were controlled using a 5dB-step custom screening hearing test made in Python 3.7.9 (Expyriment version 0.10.0). Data from nine participants were excluded because of noisy EEG recordings (high impedances or excessive eye/movement artifacts) or track loss for long temporal windows in the eye-tracking dataset, resulting in 39 participants in the final dataset.

Stimuli

We presented 6 audiovisual stimuli of ~4.5 minutes each, showing a man and a woman having a conversation. The themes of the conversations differed and were joyful to maintain attention during the whole duration of the dialogue. Dialogues contained on average 59

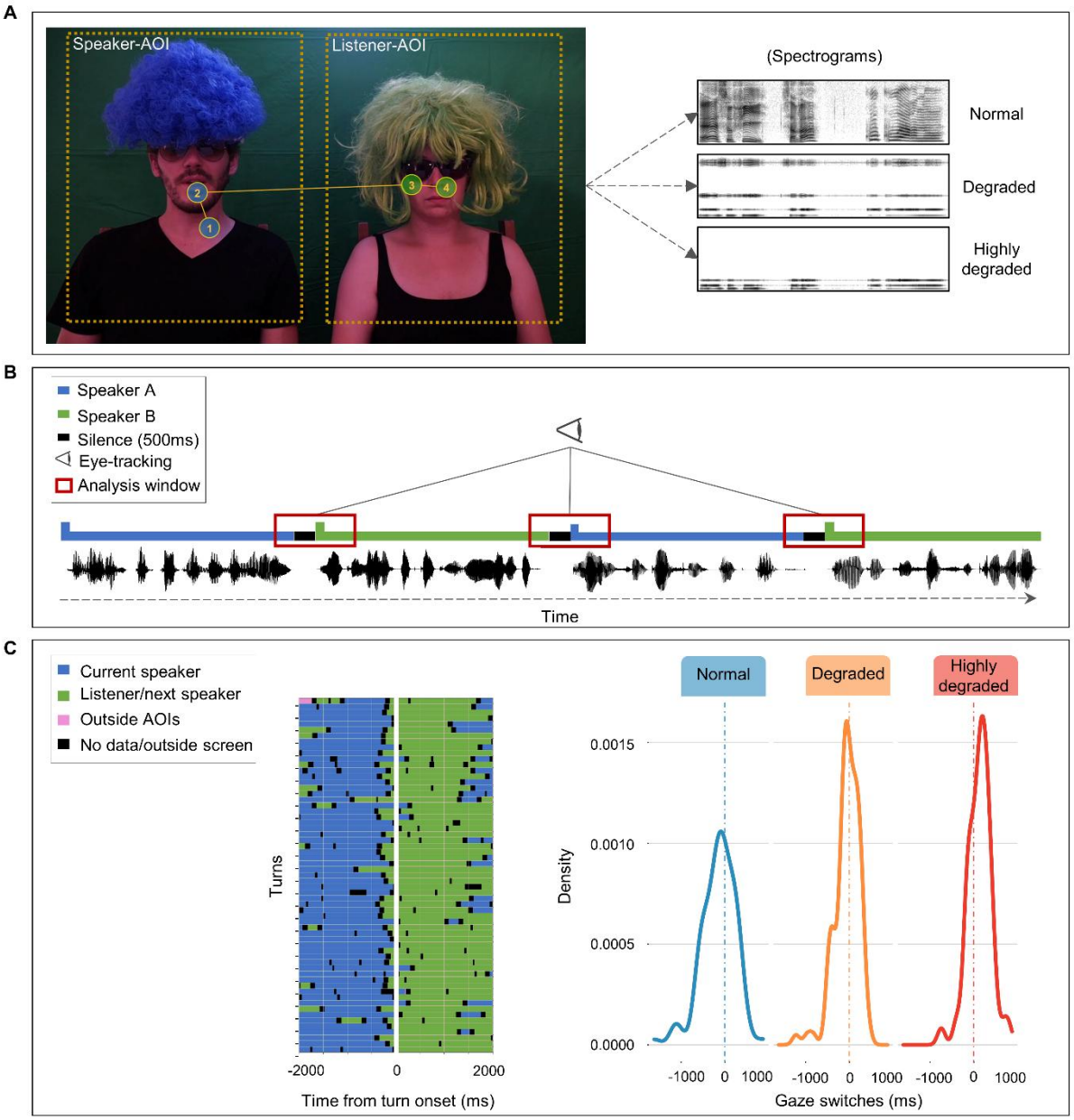
turns (min = 46; max = 66; see details in table 1). Turns lasted on average 3.75s (SD = 2.75) and global speech rate was approximately 5Hz. 60% of these turns were questions/responses adjacency pairs (min = 56 %; max = 66%) in order to elicit predictive behaviours as suggested in Casillas & Frank, 2017 (Casillas & Frank, 2017). We used Final Cut Pro X to set the gap duration at speakers' turn to 500ms in order to have homogeneous turn conditions, allow anticipatory gaze behaviour (cf. Foulsham et al. 2010; Casillas and Frank 2017) and obtain equivalent EEG baseline for turn onset analyses.

For the clarity of the evoked potentials, each turn onset began with a plosive consonant (although we avoided bilabial to avoid visual cues preceding the plosion). In the stimuli recording session, all dialogues were written and displayed on a desk in front of the speakers during video shooting. Dialogues were all read by the same male and female speaker and videos were recorded in an anechoic room. To avoid body movement when reading the dialogues, the speakers sat on a chair and their head was maintained still by a hard hat fixed to the chair. Each speaker wore sunglasses to hide eye movements and a wig to hide the hard hat (Figure 1A left). Furthermore, in order to discard spurious visual cues, the non-speaking speaker was "frozen" via video editing whenever a head or lip movement occurred before the turn. This happened in approximately ~20% of the turns. In order to avoid sharp jumps in the images of the upcoming speaker such freezing took advantage of fading and morphing technics (mMorphCut plug-in ; FINAL CUT pro X).

In other words, turns could not be anticipated on the basis of visual information of the next speaker right before the turn. Thus, anticipatory gaze behaviour cannot be interpreted in response to visual cues but as auditory predictive behaviour only. We also created 12 open questions for each dialogue (e.g. "why was Tom walking in the wood?") (see table 2 for more

examples for one dialogue). These questions were recorded using an auditory only format, by a female speaker different from the one in the dialogues.

Speech recordings of the six dialogues were degraded using a noise vocoding approach (Shannon et al. 1995; custom Matlab script). This allows to parametrically vary the spectral detail, with increasing numbers of channels associated with increasing perceptual clarity (Figure 1A right). The procedure allowed creating a highly degraded condition in which we kept only frequency bands from 120Hz to 237Hz, 405Hz to 538Hz and 919Hz to 1028Hz (i.e. 3 frequency bands left with 5 ERB scales) and a moderately degraded condition in which we removed all frequency bands except from 120Hz to 237Hz, from 674Hz to 805Hz, from 2025Hz to 2112Hz and 5208Hz to 5236Hz (so 4 frequency bands left with 8 ERB scales). The overall RMS amplitude of the audio files was adjusted to be the same across all dialogues in the three conditions (normal, moderately degraded, highly degraded). We also run a pilot test to ensure that comprehension varied as a function of speech degradation and that, in the most challenging condition, participants could still understand part of the conversation. More precisely, 6 native French participants listened to audio excerpts of the dialogues and repeated all words they could recognize. The excerpts were degraded in a parametric manner in the number of channels, the bandwidth and the low frequency cut-off. The parameters were chosen in such a way to yield a condition of moderate difficulty (comprehension level between 30 and 50%) and another condition of great difficulty (comprehension level between 10 and 30%).



175

176 Figure 1. Schematic illustration of the task and analyses. A. On the left, snapshot of the video illustrating the
177 fixation points of the participant and the gaze switch. On the right, the three different speech conditions with
178 normal, moderately degraded and highly degraded speech (using a vocoding approach). B. Schematic
179 illustration of the turns in the dialogue, the controlled gap between turns (500ms) and the gaze switch window
180 of analysis. C. On the left, gaze behaviour of a single participant showing the gaze switch for every turn of a
181 single dialogue in the normal speech condition with respect to turn onset (zero). On the right, the gaze switch
182 distribution for a single participant in the three different degradation conditions. A median split is used to class
183 turns as being more or less well anticipated.

Procedure

Participants were equipped with a 64 preamplified Ag–AgCl electrodes (International 10/10 system site, BrainAmp system). The ground electrode was placed at AFz, the reference electrode at FCz, and the EEG signal sampling rate recording was 1000 Hz. Participants were comfortably seated in a Faraday sound-proof booth in front of a computer screen (24'') with a resolution of 1920 x 1080px and a refresh rate of 100Hz), at a distance of approximately 70 cm. A Gazepoint GP3 eye-tracker (sampling rate: 60 Hz; accuracy: 0.5-1° of visual angle) was installed at the bottom of the screen on a tripod to record the participant's gaze positions. OpenSesame software (Mathôt et al., 2012) installed on a Dell laptop (Precision T1700) launched the eye-tracker recording through the PyGaze package (Dalmaijer et al., 2014) and the audio-visual stimulation. The videos were displayed on full screen resolution, and the sound was delivered through a 2040 YAMAHA amplifier and two NS 1020 Studio YAMAHA loudspeakers located on both sides of the screen. Before each video presentation, the participant's eyes were calibrated using the PyGaze's standard 12-points calibration procedure.

The experimenter asked participants to attentively follow the dialogues and informed them that, after each dialogue, they should answer 12 questions on the dialogue content. These questions allowed us to ascertain participants' attention and assess the level of comprehension. Each participant saw 2 dialogues in each of the three experimental conditions (normal speech, moderately degraded speech, highly degraded speech). The level of degradation of each dialogue and the order of the dialogues were counterbalanced across participants (each dialogue was presented at a different degradation level - normal, moderately degraded, highly degraded - to different participants).

Stimuli, eye-tracking and EEG data synchronization

A custom-made sync box based on an Arduino micro-controller ensured the synchronization between the EEG data and the acoustic stimulation that was embedded in the video file. The stereo sound goes from the stimulation PC to the sync box (audio cable). There, the stereo is split and the first channel containing the speech signals of the dialogues goes directly to the loudspeakers. The second channel signal goes through the micro-controller. This channel contains an audio trigger indicating the beginning and the end of each dialogue. The Arduino detects the audio trigger and sends an adapted signal to the EEG system. To ensure a precise synchronization of the eye-tracking on the acoustic stimulation and the EEG data, we used saccadic movements. More precisely, saccades can be easily detected in the eye-tracking data, but they are also clearly visible in the EEG signal, especially using Independant Component Analysis decomposition (Makeig et al., 1996). Thus, for each dialogue, we detected gaze switch for every turn in both the eye-tracking and EEG data (see supplementary material 1). Then, we computed for each subject and dialogue the median delay between the two signals across all turns and used it to temporally realign the eye-tracking to EEG (median delay across participants = 23ms; median sd across participants = 12ms). This procedure ensured a good time-alignment of the three types of data (video, eye-tracking and EEG).

Data Analysis

Eye-tracking

In this study we used the gaze switch from one speaker to the other as a proxy of anticipatory behaviour with respect to the upcoming turn onset (see Figure 1A left and B). However, in contrast with previous studies (Casillas & Frank, 2017) that were mostly interested in anticipatory gaze shift, we were also interested in somewhat late gaze shifts.

We assumed that adult participants need approximately 200ms to plan an eye movement (Kamide et al., 2003). Thus, we considered a temporal window around the turn allowing to keep both anticipatory and non-anticipatory gaze switch, resulting in a -1/+1s window around the turn onset (Figure 1B). Note, nonetheless that, as expected, most shifts occurred in the inter-turn gap, that is before the turn onset (Foulsham et al., 2010; Keitel et al. 2013; Casillas and Frank 2017, see Figure 2B).

Moreover, we also applied three supplementary criteria to filter spurious gaze switches. First, gaze before and after the switch should fall within an area of interest (AOIs), defined as stationary rectangle surrounding each face (see Figure 1A left). Second, gaze switch should be preceded by at least 100ms fixation on the current speaker. Third, it should be followed by at least 150ms fixation to the next speaker. Concerning the possible back and forth gaze behaviour preceding turn (4.5% in our data), we kept only the first gaze switch.

In short, we computed a gaze switch latency, that is relative to the upcoming turn onset time. A positive indicates thus a gaze switch following the turn onset, while a negative value indicates a gaze switch before the turn onset (Figure 1C left). Then, we used a median split to class turns as early or late, separately for each subject and condition (Figure 1C right).

EEG

Signal processing was done using EEGLAB (Delorme & Makeig, 2004) and custom scripts written in MATLAB. We high-pass filtered (0.5Hz) continuous data and rejected major artifacts (> 300 µV). For every participant, we systematically used Independent Component Analysis to remove eye blinks and saccadic movements and, when needed, muscular activity. We then low-pass filtered (45Hz) and segmented continuous data into epochs of 900ms starting at 100ms prior to turn onset. We zero-mean normalized epochs to the baseline ([–

100, 0]ms) and re-referenced to the algebraic average of all electrodes. Finally, we averaged epochs according to the three conditions, also separating epochs according to the latency of the gaze shift corresponding to each turn (early or late, see above).

Statistical analysis

We computed all statistical analysis using R (Team, 2021) and the lme4 package (Bates et al., 2015). We evaluated participants' comprehension scores (1 point attributed to each good answer; range 0 - 12) as a function of speech degradation by fitting linear models on the mean of correct responses for each condition (2 dialogues by condition) and each subject. Similarly, we modelled the mean of gaze switch latencies for each dialogue as a function of speech degradation by fitting linear model on latency data. We then used a general linear model (glm) in modelling the time spent gazing at the current speaker (ratio of the total fixation time). All theses models were tested against the null model before being interpreted. We computed a linear mixed model with interaction to explain comprehension score according to gaze switch latencies and speech degradation, with subject as random effect (comprehension ~ gaze switch latency * speech degradation + 1|subject). This model was compared 1) to the null model, 2) to the model with only gaze switch latencies as predictor and 3) to the additive model (gaze switch latencies and speech degradation as independent predictors). Statistical significance of the fixed effects was assessed by model comparison using the Akaike Information Criterion, thus arbitrating between complexity and explanatory power of the models. Normality and homoscedasticity of the residuals of all the models were systematically visually inspected. Reported p values are Satterthwhaite approximations obtained with the lmerTest package (Kuznetsova et al., 2017).

The statistical significance of the differences between conditions for the ERP data was evaluated by a cluster-based random permutation approach for the full set of 64 electrodes (2-tailed test with 500 permutations over the whole time range of the ERP epoch, i.e. between -100 and 800ms). This statistical approach handles the multiple-comparisons problem. More precisely, the approach controls for the Type-1 error rate in multiple testing across channels and time points by identifying clusters of significant differences between conditions in the time and space dimensions (Maris & Oostenveld, 2007). Finally, in order to assess the relation between gaze switch and ERPs, we used simple linear models with the amplitude difference between late and early trials in the N400 window explained as a function of the interaction between gaze switch latencies and speech degradation (N400 effect ~ gaze switch latency * speech degradation). As for the other analyses, this model was compared to the additive model.

Results

As expected, the comprehension score decreases as a function of increasing speech degradation (β = -2.551, SE = 0.22, t = -11.55, p < 0.001, Figure 2A). Compared to the normal condition, both the highly degraded and degraded speech conditions altered the comprehension score (β = -1.23, SE = 0.42, t = -2.929 p = 0.004 ; β = -5.102, SE = 0.42, t = -12.145, p < 0.001, respectively).

It is important to note that participants' score are well above zero and this is also the case in the highly degraded condition (range: 1-10, median 5.5). Considering that the score is based on response to open questions, it is extremely unlikely that a single correct answer is given

by chance. As such, these results show that participants of the current experiment can partly understand the dialogues even in the most difficult condition of speech perception.

When turning to eye-movements, participants gaze data are available on average for 87% of the total duration of the dialogues (range: [82-90]). The loss of data is possibly due to a gaze off the screen and/or to a temporary loss of the eye from the tracker. From these available data, it appears that participants gazed at the current speaker most of the time during video playback (93%; range: [87-97]). Compared to the normal condition, time spent on current speaker is greater in the degraded (β = 0.580, SE = 0.073, t = 7.881, p < 0.001) and highly degraded conditions (β = 0.616, SE = 0.074, t = 8.317, p < 0.001). Importantly, not only participants gazed most of the time at the current speaker, but when considering gaze switch towards the current speaker, these took place most often in a -1/+1sec window around the turn onset (mean 80 %, range: [0.56-0.96]), the rest being gaze switches far away from the turn or missing data.

Figure 2

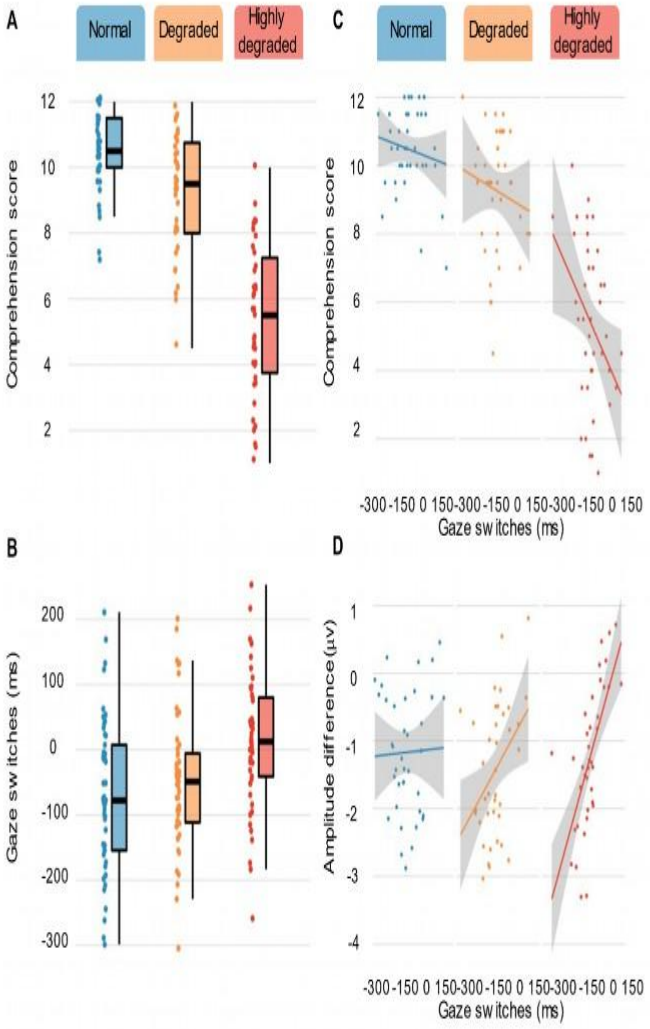


Figure 2. Behavioural results and correlations. A. Effect of speech degradation level on comprehension. Dots represent the average comprehension score for each participant in each condition. B. Effect of speech degradation level on gaze switch latency. Latency at zero milliseconds corresponds to the onset of the turn. Dots represent the average latency for each participant in each condition. C. Scatter plots and linear regression between comprehension scores and gaze switch latency in the three different conditions. D. Scatter plots and linear regression between N400 effect size (late minus early, in the 300-550ms latency window) and gaze switch latency in the three different conditions.

Normalizing by time unit this gives a number of gaze switch per second of 0.47 during the turn window and 0.08 during the rest of the dialogue (that is switches from one AOI to the

other that were performed outside the -1/+1s window around the turn, see supplementary material 2). Overall, these results clearly show that gaze and gaze switches are not randomly distributed across the dialogues. On the contrary, gaze behaviour to current speakers demonstrates that participants are tracking turns during the videos, as previously reported (Casillas & Frank, 2017). When moving to gaze switch latencies, one can see that these are influenced by speech degradation (β = 43.382, SE = 9.52, t = 4.553, p < 0.001, see Figure 2B). This is due to longer latencies to turns with highly degraded speech compared to normal and degraded speech (always p < 0.001) while there are no latency differences between degraded and normal conditions (p > 0.1). Moreover, latencies globally move from negative values in the normal condition to positive values in the highly degraded condition. Because latencies are estimated with respect to turn onset, this means that while in the normal condition participants realize the gaze switch before the turn, switches mostly take place after the turn onset in the highly degraded condition.

When looking at the relation between comprehension scores and gaze latency, this changes as a function of speech degradation (β = - 5.740, SE = 0.001, t = -3.936, p < 0.001). More precisely, in the highly degraded condition shorter gaze switch latencies to turn onsets are associated to good comprehension scores (β = - 0.008, SE = 0.003, t = -2.989, p < 0.01). This effect was not significant in the moderately degraded nor in normal conditions (p > 0.5, Figure 2C). This model with gaze switch latencies and speech degradation as interaction factors was the best compared to the additive model (AIC = 1038.5 and 1043.6 respectively ; p < 0.01 and p < 0.001).

The classification of turns on the basis of the latency of gaze switches (late vs early) shows a clear negative component for turns with late switches evolving between 350 and 700ms

post-stimulus onset over a fronto-central region (Figure 3B). This effect is visible in the three different levels of speech degradation ($p < 0.05$, FDR corrected). However, this effect seems to differ across the three levels of speech degradation (Figure 3A) insofar as the clusters start at different latencies. Because it is not appropriate to make an inferential claim on the time extension of the cluster (Sassenhagen et al., 2019), we run a further analysis using a jackknife approach in five subsequent 50ms latency windows from 150 to 400ms. This approach is considered appropriate to make inferences on the latency of an ERP effect (Miller et al., 1998). Results show that differences between early and late turns are significant starting 200ms post turn onset in the normal condition, 150ms in the moderately degraded condition and 300ms in the highly degraded condition ($p < 0.05$, FDR corrected).

Finally, when looking at the relation between the ERP effect (late vs early gaze switches) in the 300-550ms window and the global gaze switch latency of participants, we find a significant interaction between gaze switch latencies and speech degradation levels ($F = 7.135$, $p = 0.001$). More precisely, one can notice, that, in the highly degraded condition, participants with overall shorter gaze switch latencies to turn onsets show a larger effect ($\beta = 0.007$, SE = 0.001, $t = 3.753$, $p < 0.001$). A similar but not significant trend is also visible in the moderately degraded ($\beta = 0.003$, SE = 0.001, $t = 1.912$, $p = 0.058$) but not in the normal

conditions (β = 0.000, SE = 0.001, t = 0.204, p = 0.838, Figure 2D).

Figure 3

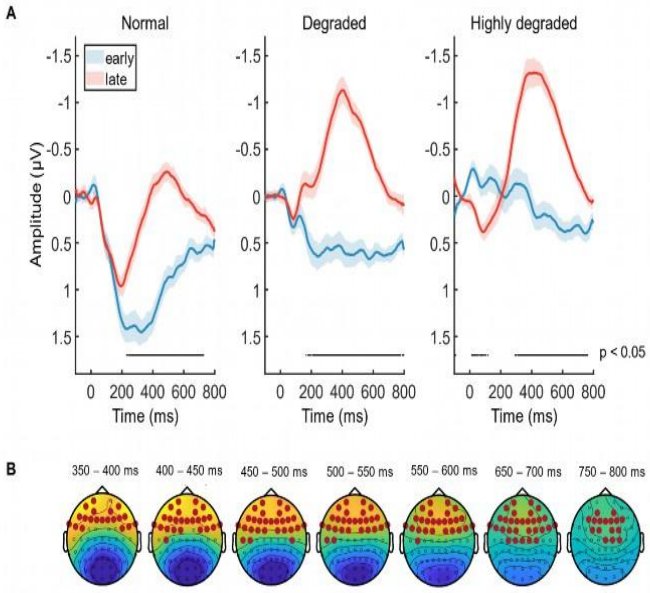


Figure 3. A. event-related potentials (ERPs) in a fronto-central region of interest time-locked to the turn onset for early and late gaze switch latencies. Shaded areas indicate the standard error of the mean. The black line below the ERPs indicates the FDR corrected significant difference between early and late conditions. B. Cluster-based statistical analyses on all electrodes showing the main effect of gaze behaviour (early vs late). The topography illustrates the difference between late and early turns across all levels of degradation. The red dots represent electrodes with significant differences (cluster corrected).

Discussion

In the present experiment, participants watched a series of videos showing two interlocutors having a conversation. We parametrically degraded speech and measured eye-movements, EEG and dialogue comprehension to gather a deeper understanding of the mechanisms underlying speech comprehension under challenging conditions. We build on the fact that

participants switch gaze at the current speaker and use this measure as a proxy of prediction to class dialogue turns as more or less well predicted and study ERPs accordingly. Results show that 1) Participants switch gaze towards the current speaker in a limited temporal window around the turn. 2) Speech degradation reduces the anticipatory gaze behaviour. 3) Globally, participants with a low comprehension score tend to have a later gaze behaviour when speech is strongly degraded. 4) ERPs to turn onsets vary as a function of the latency of the gaze switch. 5) Finally, participants with overall shorter switch gaze latencies show a better comprehension score and a stronger anticipatory effect in their neural responses; this relation between behavioural and neural variables is significant only in the most challenging (degraded) condition. We discuss these findings with respect to the relation between predictions, neural predictions errors and speech comprehension under challenging conditions. Of course, these findings are limited by the constraints inherent to the present experiment, wherein predictions are possible purposely on the basis of auditory cues only because we controlled all extralinguistic visual cues (e.g. lip preparatory movement) that undoubtedly play an important role in conversation.

Our study replicates previous findings (Casillas & Frank, 2017; Foulsham et al., 2010; Keitel et al., 2013), showing that gaze switch takes place in a precise temporal window centered before turn onset. Considering the short period between turns and the time needed to prepare and realize a saccade, gaze switch, in this context, can be interpreted as anticipatory behaviour (Keitel & Daum, 2015). The new result, in this respect, is that gaze switch behaviour is affected by the more or less challenging speech comprehension, here obtained by spectrally degrading speech signal. The reduction of anticipatory gaze behaviour in presence of stronger speech degradation seems to indicate that increasing the cost of speech processing has a detrimental effect on the prediction of the upcoming turn. This is in

399 line with results on hearing-impaired individuals showing a delay (greater effort) in
400 processing sentences with more or less contextual cues, compared to normal-hearing (Winn,
401 2016).

402 Importantly, our results show a relation between comprehension score and turn anticipatory
403 behaviour, under challenging conditions. This result is not trivial. Indeed, the comprehension
404 score reflects the global understanding of the whole dialogues, because it corresponds to
405 questions that were asked at the end of each dialogue. By contrast, gaze switch and EEG
406 dynamics reflect anticipatory behaviour at precise temporal windows surrounding
407 conversational turns. This means, that, under challenging conditions, participants that are
408 able to understand can take advantage of an anticipatory gaze behaviour. Alternatively,
409 participants with an anticipatory gaze behaviour are better able to understand. While the
410 present design does not allow for a causal interpretation between gaze behaviour and global
411 understanding of the dialogues, we will see that ERP results play here a very important role.
412 However, before discussing this point one has to interpret the ERP components, in particular
413 the negative component peaking around 400ms.

414 First, one may raise the possibility that residual gaze-related artefact may still be present in
415 the EEG, even following ICA procedures (see methods). However, 1) the systematic
416 difference in latency between late and early switches was on average 480ms, while we do
417 not observe such a delay in ERPs to early and late turns; 2) while highly degraded speech
418 induces overall later gaze switches, the N400 peak in this condition has the same latency as
419 in the other conditions (in Figure 3A the N400 peak is always 400ms). Thus, overall, one can
420 consider that differences in ERPs are due to cognitive processes and not to residual
421 physiological artefacts due to eye movements.

The first and most likely interpretation of the larger negativity to both degraded turns and late turns is in terms of an N400 like component. Indeed, the amplitude of the N400 is strongly affected by expectations: the more a target word is unexpected within a sentence context, the larger the N400 amplitude (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984). While the N400 topography we report is more anterior compared to the classic centro-parietal distribution, this may be due to the audiovisual modality (Kelly et al., 2004) or to the presence of complex scenes (Kutas & Federmeier, 2011) inducing a topography more frontal than for written words. Importantly, the N400 has been proposed to be a proxy of prediction errors (Bornkessel-Schlesewsky & Schlesewsky, 2019; Rabovsky & McRae, 2014), that play a key role within the predictive coding framework (Friston, 2010; Schultz & Dickinson, 2000).

As stated above, in order to assess the relevance of predictive coding theories with respect to speech comprehension, it is necessary to have access to both prediction and prediction error indicators. Our results, as well as previous studies, seem quite clear in showing that gaze switch behaviour can be considered as a proxy of predictions. Previous research using an electrophysiological marker of prediction (the semantic prediction potential, SPP) showed that smaller SPP amplitude (weak predictions) result in a large N400 (Grisoni et al., 2021). Similarly, our results show that turns preceded by a late gaze switch, indicating poor predictions, give rise to a larger N400 like complex compared to turns preceded by an early gaze switch. An interesting advantage of the current design is that the target words (here turn onsets) do not need to differ in content or in context in order to engender a N400. This is different for instance, from previous studies manipulating the semantic and temporal turn relation to induce an N400 effect (Bögels et al., 2015). In other words, in the current study, the word and the context eliciting or not the N400 are strictly identical and we take advantage of the natural variability in the accuracy of participants' predictions across turns.

The interpretation in terms of N400 also fits quite well with the literature on weaker expectancies under degraded speech. Previous work showed that adverse listening conditions narrow the expectancies about the upcoming speech. This is visible in a reduced N400 effect in response to incongruent or less likely words under acoustic degradation (Aydelott et al., 2006; Obleser & Kotz, 2011; Strauß et al., 2013). The present findings show a somewhat similar result, with an N400 effect that is significant in a later latency window in the most challenging condition (strong speech degradation) compared to the two other conditions. While the typical N400 effect concerns the amplitude of the negative wave, several works reported later N400 latencies in context presenting more difficult semantic access (Deacon et al., 1995; Moreno & Kutas, 2005). Interestingly, the temporal window analysis shows that traces to early and late turns diverge first in the degraded condition, then in the normal and last in the highly degraded conditions. It thus seems as if the advantage of an early gaze switch was integrated faster in moderately challenging conditions. There, the benefit of an anticipatory gaze behaviour and temporally precise (local) predictions may be maximal compared to the two other conditions wherein uncertainty may either be too low (normal speech) or too high (highly degraded speech).

However, while the N400 latency is earliest in the degraded condition (Figure 3A), the N400 amplitude best correlates with turn anticipatory behaviour in the highly degraded condition (Figure 2D). In considering this inconsistency one should keep in mind that 1) the N400 effect, in terms of amplitude, is also significant in the highly degraded condition; 2) while the correlation between N400 amplitude and average gaze behaviour does not reach the significance threshold (p = .06), the trend of the correlation is similar across the two degraded conditions.

Importantly, Figure 2C and 2D clearly show that there is a similar relation between gaze behaviour and the semantic access of dialogues. Differently from global comprehension score, the N400 amplitude does not reflect a global understanding of the conversation, but it is a rather local measure, as it is the case with gaze switch behaviour. Moreover, the two measures are serially ordered in time, with gaze switch taking place earlier ~400ms before the N400 peak. Thus, one can hypothesize that participants with an overall early gaze switch, by having a better prediction of the turn, have an easier lexical-semantic integration of the word starting the turn in the conversational context. This relation between gaze and both local (N400 effect) and global (comprehension scores) understanding is only shown under challenging conditions, wherein indeed participants can benefit of a predictive behaviour to improve their understanding of the dialogues.

It seems appropriate here to evoke an alternative (although not necessarily the most likely) interpretation of the negative component in terms of a phonological mismatch negativity (pMMN). Several studies have addressed the timing of audio-visual integration. For instance, in McGurk illusion, the temporal window allowing modality fusion and allusion ranges between 30ms and 170ms of asynchrony (van Wassenhove et al., 2007). When looking at the electrophysiological response to the incongruent audio-visual stimulation, several authors report a pMMN (Colin et al., 2002; Eskelund et al., 2015; Stekelenburg & Vroomen, 2012). The latency of this response is later than the classic MMN and is described approximately 500ms after the voice onset (Proverbio et al., 2018). In our study, the classification of turns as being accompanied by late or early gaze switch implies that the audio-visual relation of the stimuli differs in the two classes. For early gaze switch, the gaze is well anticipated and thus the audio-visual information is perceived simultaneously. By contrast, in turns with a late gaze switch, often occurring after the voice onset, the

audiovisual integration has a higher level of uncertainty. In other words, if a gaze switch takes place later than a turn onset, there may be an audiovisual mismatch because the participant is looking at the previous speaker while listening to the upcoming speaker's voice. In this context, looking at the closed lips of the previous speaker and hearing the word pronounced by the following speaker, may engender a phonological audiovisual MMN.

We would like to end the discussion with a consideration on the implicit and explicit role of predictions. It is known that the influences of the visual system on auditory perception can be so strong as to override under certain conditions the original input of the auditory system (Mcgurk & Macdonald, 1976). More commonly, viewing the lips provides relevant complementary information and can augment and improve auditory capabilities (Calvert et al., 1997; Drijvers & Özyürek, 2017; Grant & Seitz, 2000; Sumby & Pollack, 1954). In this context, it is interesting to note that, compared to the normal condition, the moderately degraded condition yields a poorer comprehension but similar gaze switch latencies and an earlier N400 effect. These preserved latencies, in a context of poorer comprehension that should rather induce poorer predictions, may be mediated by an active compensatory strategy, explicitly making greater use of visual cues to improve predictive processes (Sohoglu & Davis, 2016). Such a compensatory strategy may not hold when speech is highly degraded, yielding to slower gaze switches due to a poor understanding of the semantic and syntactic context. The respective role of implicit and explicit predictions and their link to conversation strategies (Hadley et al., 2021) will require further work addressing this specific question. Another interesting aspect that will require further work is the inter-relation between local predictions (gaze), local semantic processing (N400) and global comprehension (here assessed via specific questions on the dialogues). While on one side gaze switches temporally precede auditory and semantic processing of the upcoming turn,

this anticipatory behaviour can only build on the global comprehension of the dialogue, that, in turn, depends on the integration of local semantic processing.

To conclude, this study confirms that gaze switch can be used as a proxy of predictions in a conversational context. It shows that these predictions are related to the lexico-semantic processing of the turn start, as estimated by a neural marker of prediction error. It also shows that predictions are less accurate in more challenging listening conditions, but that they are also most useful in that specific context to make sense of the upcoming turn and, more generally, are a good indicator of global dialogue comprehension.

Acknowledgements

We would like to thank Laura Leone, Romane Pradels and Benjamin Morillon for their participation in the design of the dialogues and the production of videos.

Funding sources: Work supported by APA foundation (RD-2016-9), ANR-16-CONV-0002 (ILCB) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

References

Aydelott, J., Dick, F., & Mills, D. L. (2006). Effects of acoustic distortion and semantic context on event-related potentials to spoken words. *Psychophysiology*, *43*(5), 454–464. https://doi.org/10.1111/j.1469-8986.2006.00448.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beňuš, Š., Gravano, A., & Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, *43*(12), 3001–3027. https://doi.org/10.1016/j.pragma.2011.05.011

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never say no . . . How the brain interprets the pregnant pause in conversation. *PLoS ONE*, *10*(12). https://doi.org/10.1371/journal.pone.0145474

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology*, *10*(FEB), 298. https://doi.org/10.3389/fpsyg.2019.00298

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596. https://doi.org/10.1126/science.276.5312.593

Casillas, M., & Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *Journal of Memory and Language*, *92*, 234–253. https://doi.org/10.1016/j.jml.2016.06.013

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology*, *113*(4), 495–506.

https://doi.org/10.1016/S1388-2457(02)00024-X

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2014). PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods*, *46*(4), 913–921. https://doi.org/10.3758/s13428-013-0422-2

De Ruiter, J., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language*, *82*, 515–535. http://dx.doi.org/10.1353/lan.2006.0130

Deacon, D., Mehta, A., Tinsley, C., & Nousak, J. M. (1995). Variation in the latencies and amplitudes of N400 and NA as a function of semantic priming. *Psychophysiology*, *32*(6), 560–570. http://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=11064097&lang=fr&site=ehost-live

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, *13*, 9–21. https://doi.org/http://dx.doi.org/10.1016/j.jneumeth.2003.10.009

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Eskelund, K., MacDonald, E. N., & Andersen, T. S. (2015). Face configuration affects speech perception: Evidence from a McGurk mismatch negativity study. *Neuropsychologia*, *66*, 48–54. https://doi.org/10.1016/j.neuropsychologia.2014.10.021

582 Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a
583 dynamic situation: Effects of social status and speaking. Cognition, 117(3), 319–331.
584 https://doi.org/10.1016/j.cognition.2010.09.003

585 Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal*
586 *Society B: Biological Sciences*, *360*(1456), 815–836.
587 https://doi.org/10.1098/rstb.2005.1622

588 Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews*
589 *Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

590 Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory
591 detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3),
592 1197. https://doi.org/10.1121/1.1288668

593 Grisoni, L., Tomasello, R., & Pulvermüller, F. (2021). Correlated Brain Indexes of Semantic
594 Prediction and Prediction Error: Brain Localization and Category Specificity. *Cerebral*
595 *Cortex*, *31*(3), 1553–1568. https://doi.org/10.1093/cercor/bhaa308

596 Hadley, L. V., Whitmer, W. M., Brimijoin, W. O., & Naylor, G. (2021). Conversation in small
597 groups: Speaking and listening strategies depend on the complexities of the
598 environment and group. Psychonomic Bulletin & Review, 28(2), 632-640.

599 Hirvenkari, L., Ruusuvuori, J., Saarinen, V.-M., Kivioja, M., Peräkylä, A., & Hari, R. (2013).
600 Influence of Turn-Taking in a Two-Person Conversation on the Gaze of a Viewer. *PLoS*
601 *ONE*, *8*(8), e71569. https://doi.org/10.1371/journal.pone.0071569

602 Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in
603 incremental sentence processing: Evidence from anticipatory eye movements. *Journal*

604 *of Memory and Language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-
605 596X(03)00023-8

606 Keitel, A., & Daum, M. M. (2015). The use of intonation for turn anticipation in observed
607 conversations without visual signals as source of information. *Frontiers in Psychology*,
608 *6*(February), 108. https://doi.org/10.3389/fpsyg.2015.00108

609 Keitel, A., Prinz, W., Friederici, A. D., Hofsten, C. von, & Daum, M. M. (2013). Perception of
610 conversations: The importance of semantics and intonation in children's development.
611 *Journal of Experimental Child Psychology*, *116*(2), 264–277.
612 https://doi.org/10.1016/j.jecp.2013.06.005

613 Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the
614 N400 component of the event-related brain potential (ERP). *Annual Review of*
615 *Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

616 Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy
617 and semantic association. *Nature*, *307*(5947), 161–163.

618 Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in
619 Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.
620 https://doi.org/10.18637/jss.v082.i13

621 Levinson, S. C. (2016). Turn-taking in Human Communication - Origins and Implications for
622 Language Processing. *Trends in Cognitive Sciences*, *20*(1), 6–14.
623 https://doi.org/10.1016/j.tics.2015.10.010

624 Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent component analysis
625 of electroencephalographic data. *Advances in Neural Information Processing Systems*,

145–151.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Miller, J., Patterson, T. U. I., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. Psychophysiology, 35(1), 99-115.

Molinaro, N., Monsalve, I. F., & Lizarazu, M. (2016). Is there a common oscillatory brain mechanism for producing and predicting language? *Language, Cognition and Neuroscience*, *31*(1), 145–158. https://doi.org/10.1080/23273798.2015.1077978

Moreno, E. M., & Kutas, M. (2005). Processing semantic anomalies in two languages: An electrophysiological exploration in both languages of Spanish-English bilinguals. *Cognitive Brain Research*, *22*(2), 205–220. https://doi.org/10.1016/j.cogbrainres.2004.08.010

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage*, *55*(2), 713–723. https://doi.org/10.1016/j.neuroimage.2010.12.020

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Proverbio, A. M., Raso, G., & Zani, A. (2018). Electrophysiological Indexes of Incongruent Audiovisual Phonemic Processing: Unraveling the McGurk Effect. *Neuroscience*, *385*, 215–226. https://doi.org/10.1016/j.neuroscience.2018.06.021

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*(1), 68–89. https://doi.org/10.1016/j.cognition.2014.03.010

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, *85*(3), 618–660. https://doi.org/10.1037/0033-2909.85.3.618

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, *6*, 509. https://doi.org/10.3389/fpsyg.2015.00509

https://doi.org/10.1093/cercor/bhl024

668 Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data
669 do not establish significance of effect latency or location. Psychophysiology, 56(6),
670 e13335.

671 Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of*
672 *Neuroscience*, *23*, 473–500. https://doi.org/10.1146/annurev.neuro.23.1.473

673 Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-
674 constructional units and turns in conversation. *Pragmatics. Quarterly Publication of the*
675 *International Pragmatics Association (IPrA)*, *6*(3), 371–388.
676 https://doi.org/10.1075/prag.6.3.06sel

677 Shannon, R. V, Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition
678 with primarily temporal cues. *Science*, *270*(5234), 303–304.
679 https://doi.org/10.1126/science.270.5234.303

680 Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing
681 prediction error. *Proceedings of the National Academy of Sciences of the United States*
682 *of America*, *113*(12), E1747–E1756. https://doi.org/10.1073/pnas.1523266113

683 Šoškić, A., Jovanović, V., Styles, S. J., Kappenman, E. S., & Ković, V. (2021). How to do better
684 N400 studies: reproducibility, consistency and adherence to research standards in the
685 existing literature. *Neuropsychology Review*, 1-24.

686 Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical
687 Review and Theoretical Interpretation. *Language and Linguistics Compass*, *9*(8), 311–
688 327. https://doi.org/10.1111/lnc3.12151

689 Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding

690 of auditory location in the perception of natural audiovisual events. *Frontiers in*
691 *Integrative Neuroscience*, *6*(MAY 2012), 1–7. https://doi.org/10.3389/fnint.2012.00026

692 Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G.,
693 Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural
694 variation in turn-taking in conversation. *Proceedings of the National Academy of*
695 *Sciences of the United States of America*, *106*(26), 10587–10592.
696 https://doi.org/10.1073/pnas.0903616106

697 Strauß, A., Kotz, S. A., & Obleser, J. (2013). Narrowed expectancies under degraded speech:
698 Revisiting the N400. *Journal of Cognitive Neuroscience*, *25*(8), 1383–1395.
699 https://doi.org/10.1162/jocn_a_00389

700 Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise.
701 *Journal of the Acoustical Society of America*, *26*(2), 212–215.
702 https://doi.org/10.1121/1.1907309

703 Team, R. C. (2021). *R: A language and environment for statistical computing. Vienna: R*
704 *Project, 2017*.

705 Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition
706 processes. *Perception & Psychophysics*, *34*(5), 409-420

707 van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in
708 auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607.
709 https://doi.org/10.1016/j.neuropsychologia.2006.01.001

710 Winn, M. (2016). Rapid Release From Listening Effort Resulting From Semantic Context, and
711 Effects of Spectral Degradation and Cochlear Implants. *Trends in Hearing*, *20*, 1–17.

712 https://doi.org/10.1177/2331216516669723

713 Supplementary material

Dialogue's number	Dialogue's length (seconds)	Number of turns	Number of questions/response pairs
1	269	60	40
2	247	46	26
3	185	62	36
4	258	60	38
5	253	66	40
6	248	62	36

714

715 Table 1. Dialogues's details.

716

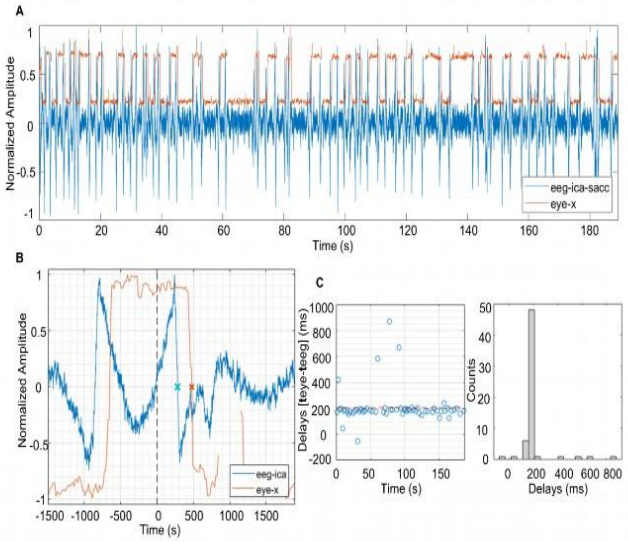
Est-ce que Damien veut attraper une bête avec sa cage ? *Does Damien want to catch an animal with his cage?*
Cette cage est faite pour qui ? *Who is this cage for?*
Comment Damien arrive à rentrer dans sa cage ? *How does Damien get into his cage?*
Dans quoi peut rentrer le génie d'Aladin ? *What can Aladdin's genie fit into?*

Pourquoi Damien veut se cacher dans une cage ? *Why does Damien want to hide in a cage?*
Comment se déplace l'étrange fée ? *How does the strange fairy move?*
Qui porte un grand chapeau ? *Who is wearing a big hat?*
Pourquoi la sorcière veut attraper Damien ? *Why does the witch want to catch Damien?*
Camille va chercher qui pour aider Damien à se débarrasser de la méchante sorcière ? *Who will Camille look for to help Damien get rid of the wicked witch?*
Dans cette histoire, où habite Harry Potter ? *In this story, where does Harry Potter live?*
Pourquoi Camille pense qu'une bête s'est échappée ? *Why does Camille think an animal has escaped?*
Est-ce que Harry Potter aime les vilaines sorcières ? *Does Harry Potter like bad witches?*

717

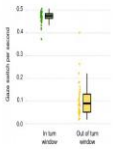
718 Table 2. Exemple of questions for one dialogue.

719



720

721 Supplementary material 1. Example of eye-tracking and EEG synchronization for one dialogue. A. Overlapping
722 of ICA-EEG data and Eye-tracking data. B. Zoom in for one trial (one turn); latency at zero milliseconds
723 corresponds to the onset of this turn. C. Dashed red line represents the mean of delays between Eye-tracking
724 and EEG data for the same subject in the same dialogue as in A.



Supplementary material 2. Number of gaze switch per second in and out of the turn window.