DISCOURSE PROCESSES
2025, VOL. 62, NO. 2, 69–88
https://doi.org/10.1080/0163853X.2024.2413314

Routledge
Taylor & Francis Group

Multimodal information density is highest in question beginnings, and early entropy is associated with fewer but longer visual signals

James P. Trujillo [a,b] and Judith Holler[a,b]

[a]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands; [b]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ABSTRACT

When engaged in spoken conversation, speakers convey meaning using both speech and visual signals, such as facial expressions and manual gestures. An important question is how information is distributed in utterances during face-to-face interaction when information from visual signals is also present. In a corpus of casual Dutch face-to-face conversations, we focus on spoken questions in particular because they occur frequently, thus constituting core building blocks of conversation. We quantified information density (i.e. lexical entropy and surprisal) and the number and relative duration of facial and manual signals. We tested whether lexical information density or the number of visual signals differed between the first and last halves of questions, as well as whether the number of visual signals occurring in the less-predictable portion of a question was associated with the lexical information density of the same portion of the question in a systematic manner. We found that information density, as well as number of visual signals, were higher in the first half of questions, and specifically lexical entropy was associated with fewer, but longer visual signals. The multimodal front-loading of questions and the complementary distribution of visual signals and high entropy words in Dutch casual face-to-face conversations may have implications for the parallel processes of utterance comprehension and response planning during turn-taking.

Introduction

Face-to-face conversation involves the rapid back-and-forth of turn-taking, whereby speakers exchange information through multimodal signals. Critically, the fast pace of turn-taking requires next speakers to process the currently unfolding multimodal utterance in parallel to planning their own response. A critical open question is whether the information in a multimodal utterance is structured in such a way as to facilitate this parallel planning and listening, particularly when the utterance is likely to receive a response.

Previous research shows that next speakers begin planning a response as soon as they have enough information to do so (Barthel & Levinson, 2020; Barthel et al., 2017; Bögels, 2020; Bögels et al., 2015; Magyari et al., 2014). Whereas this early planning allows next-speakers to launch their response within a very short time after the current speaker's utterance finishes (Corps, Crossley et al., 2018; Corps, Gambi et al., 2018; Levinson, 2016; Magyari et al., 2014, 2017), the response planning may adversely affect how well the unfolding utterance is processed after planning has started. In particular, this parallel planning has been shown to generally increase cognitive load (Barthel & Sauppe, 2019), as well

CONTACT James P. Trujillo, James.Trujillo@donders.ru.nl Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, Postbus 9104, Nijmegen 6500HE, The Netherlands

Supplemental data for this article can be accessed online at https://doi.org/10.1080/0163853X.2024.2413314

as to reduce semantic processing (Barthel, 2021) and anticipatory processing of the ongoing utterance in fast responders (Bögels et al., 2018). These findings suggest that information arriving later in an utterance may be processed less well than earlier information because next-speakers are simultaneously engaged with planning their own response.

An important question that arises from these studies of paralleling planning and comprehension is whether information is distributed in utterances in such a way as to make dual planning easier. Such an information distribution would be particularly relevant for so-called *adjacency pairs*, in which the first utterance sets normative expectations for the type of response to follow (e.g., greetings are followed by greetings, questions followed by responses; Kendrick et al., 2020; McHoul, 2008; Raymond, 2016; Schegloff, 2007; Stivers, 2013). In the case of adjacency pairs such as question–response sequences, the timing of the response carries pragmatic meaning, with longer gaps between question and response being perceived as a signal that a dispreferred response will be given (Kendrick & Torreira, 2015). It is therefore important that the next-speaker is able to plan and initiate a response in a timely manner and that this response is appropriate for the question (i.e., taking into account any and all relevant information).

Fast, sufficient comprehension may be facilitated by more unpredictable information occurring early in the utterance. Both *surprisal* and *next-word entropy* have been used to quantify the distribution of lexical *information density* (Aurnhammer & Frank, 2019; Lowder et al., 2018), and the two measures capture different, yet complementary aspects of information density and predictability. Namely, *surprisal* measures how unexpected a word is given its preceding context and is thus a backward-looking measure of informativity. Conversely, *next-word entropy* captures the uncertainty about the next word and is thus a forward-looking measure of uncertainty. To date, only surprisal has been reported for within-utterance information distribution, although next-word entropy likely provides a complementary quantification.

To understand how information is distributed in natural face-to-face conversation, however, we must also consider the multimodal nature of communication. When speaking to one another, we also use movement of the hands, such as iconic manual gestures, to clarify or add to the information in speech, as well as more pragmatic and interactive gestures to shape the stance that we wish to signal (e.g., raising the hands in front of one's chest, with palms facing forwards, to distance one's self from what is said), to structure the timing of the interaction or to provide visual emphasis to what is being said (Bavelas et al., 1992; Kendon, 1985, 2017; McNeill, 2000, 2016). We also prominently use the many face muscles, raising or furrowing eyebrows, smiling, drawing the corners of the mouth down in a "facial shrug," and other facial visual signals that signal our emotional state (Ekman, 1993, 2004; Jack & Schyns, 2015; Snoek et al., 2023) as well as conversational intentions (Bavelas et al., 2014; Domaneschi et al., 2017; Nölle et al., 2021; Nota et al., 2021, 2022). All of these visual signals thus critically contribute to the semantic and pragmatic interpretation of an utterance (Holler & Levinson, 2019; Trujillo & Holler, 2023), and thus can be seen as relevant information sources when considering information distribution.

Indeed, visual facial signals have been shown to pattern differently according to an utterance's function, or social action. The range of social actions questions can fulfill are wide ranging. For example, one can produce an utterance that is simply a statement and requires no immediate response, or one can perform a question, typically requiring a response from the addressee. Further, questions can perform even more specific social actions (Atkinson et al., 1984; Kendrick et al., 2020; Levinson, 2017), such as requesting factual information (e.g., "What time is it?"), checking mutual understanding (e.g., "Is this the same person you mentioned earlier?"), or expressing a stance or sentiment (e.g., "Isn't that very far away?"). Recent research has shown that some (or combinations of) facial signals are more likely to occur with questions, whereas other combinations are more likely to occur with responses (Nota et al., 2021), and even the more specific categories of questions are associated with different patterns of facial signals during natural conversation (Nota et al., 2023). Similarly, questions are often marked by particular manual gestures (Cooperrider et al., 2018; Mondada, 2007; Müller, 2004), and the occurrence of gestures seems to facilitate the processing of questions in

conversation (Holler et al., 2018). Therefore, visual signals contribute to the information content of utterances (Gerwing & Allison, 2009; Holler & Beattie, 2002, 2003; Hostetter, 2011; Trujillo & Holler, 2024a; Zhang et al., 2021) as well as to the processing (Drijvers & Holler, 2023) and semantic and pragmatic interpretation of spoken utterances (Domaneschi et al., 2017; Holler et al., 2011; Kelly et al., 1999; Nota et al., 2022; Özyürek et al., 2007), and thus must be considered as forming part of the overall information distribution of an utterance.

If visual signals are also contributing to the information of an utterance, then understanding the patterns of information distribution requires us to investigate how visual signals associate with the distribution of linguistic information. Specifically, although lexical informativity has been shown to be unevenly distributed between the first and second halves of an utterance in various languages (Trujillo & Holler, 2024b), it is currently not known how visual information interacts with the spoken information. While we are not aware of any empirically tested measure of visual informativity that extends to both manual gestures (especially nonrepresentational ones) and facial signals, the mere presence of visual signals may also be taken as providing information (at least as a basic measure). That is, a word with higher surprisal than the previous words is providing more information than a word with lower surprisal, whereas the presence of a visual signal (e.g., an eyebrow raise) results in an utterance with more information than if there were no visual signal (although in some cases the lack of an expected expression may also be meaningful). Although the informativity of visual signals does not currently have a measure analogous to lexical informativity, investigating whether and how the number and prominence of visual signals relates to lexical information would provide a first step in understanding how different sources of multimodal information are jointly distributed.

The present study aims to address the question of multimodal information distribution by utilizing a large corpus of face-to-face, conversational interactions between Dutch acquaintances, and analyzing lexical information density and the occurrence and (temporal) prominence of facial and manual gestures. Some languages have been shown to have an uneven distribution of information, falling in two groups, with German, Japanese, and Egyptian Arabic carrying more unpredictable information in the first half of the utterance (Trujillo & Holler, 2024b), and English, Spanish, and Mandarin showing the opposite pattern (Klafka & Yurovsky, 2021; Trujillo & Holler, 2024b). More specifically, in Trujillo and Holler (2024b), the pattern of front- versus back-loaded information is suggested to be related to the distribution of nouns and verbs within utterances. In some languages, nouns and verbs naturally occur more frequently in the first half of utterances, and in these same languages the first half of the utterance carries more unpredictable information (i.e., is more informative) than the last half. In other languages, this pattern is reversed. These different patterns of information distribution may be related to more general language preferences, such as whether short (English) or long (Japanese) phrases preferentially occur earlier within an utterance, or whether more important or urgent information preferentially occurs earlier or later in an utterance (Levshina et al., 2023). For example, along phrase may be an argument with along series of modifying words, which typically are produced first in Japanese (Yamashita & Chang, 2001), afront-loaded language, but are produced later in English (Stallings et al., 1998), aback-loaded language. However, this study did not consider the occurrence of visual signals. Therefore, we considered the larger range of information sources that language use actually entails based on this expectation of differences in information distribution between the first versus last half of speaking turns.

Specifically, we quantified lexical information density both in terms of *surprisal* (i.e., *informativity*), and *next-word entropy* (i.e., *uncertainty*), and annotated a range of manual and facial signals produced during speech. We then tested the following: 1) whether Dutch utterances showed a difference in surprisal and/or entropy, as well as number of visual signals, between the first and last portion of an utterance; and 2) whether the lower-predictability portion of the utterance was associated with more or less visual signaling compared with the higher-predictability portion, also looking at the specific visual signals (i.e., movement of particular articulators, such as the eyebrows or hands) that may be driving this effect. To address the fact that visual informativeness can relate to both the number of unique signals, or their duration, or both, we additionally tested for: 3) an association between number

and relative duration of visual signals. This approach provides a more complete picture of the extent of visual signaling and how it relates to lexical informativeness. Duration is used in conjunction with number of signals because previous research suggests that the duration of visual signals (or at least manual gestures) is associated with their informativeness. Specifically, manual gestures are shorter in contexts with more common ground between speakers compared with when there is more common ground (Holler et al., 2022).

Regarding differences in lexical information density between first and last utterance halves, one hypothesis is that first halves will show higher information density, in line with earlier parallel planning-listening results (e.g., Barthel & Levinson, 2020; Bögels et al., 2015). An alternative hypothesis is that higher information density will be found in the last half, which may occur despite parallel planning–listening and be more related to traditional notions of linguistic information structuring preferences, such as providing given (low-uncertainty/entropy) before new (high-uncertainty /entropy) information (Halliday, 1967).

Regarding the association between lexical information density and visual signaling, one hypothesis is that higher lexical information density (i.e., high surprisal or entropy) would be associated with lower visual signaling because this density also leads to a complementary distribution of information overall (i.e., high lexical informativeness but low visual informativeness). An alternative possibility is that visual signals pattern in a parallel fashion to lexical information density, with more visual signaling occurring in the parts of turns with higher information density (i.e., high lexical informativeness and high visual informativeness).

As an additional exploratory analysis, we also draw on extant data about the intention (or social action) conveyed by the question turns we analyzed, although we had no specific hypothesis about the relation between social action and information distribution. because this part of the analysis was less central to our undertaking, we report the results in the Appendix.

Overall, the study will provide novel insights about how multimodal information is structured in natural conversation, which will inform models of situated language production. The current work marks a step forward from studies that focus on unimodal language use only, and it moves us closer to seeing "the whole picture" of how communicative behaviors and information are coordinated during face-to-face conversation.

Methods

Materials

Data for the present study consisted of video and audio data collected from 34 acquainted dyads conversing (68 participants, 51 female, mean age 23.10 years). Participants were recruited via the participant database of the Max Planck Institute for Psycholinguistics. Each participant thus recruited was asked to attend the session with a friend. The length of time for which participants had been friends varied across dyads from a few weeks to many years, as did their indication of how strong their relationship was. Each dyad engaged in three 20-minute Dutch conversations: first in an entirely unprompted, casual conversation, then in a conversation about three pre-defined topics (see Appendix I for the original topics and instructions), and third in a conversation that focused on a collaborative planning task. Recordings were made in a soundproof room at the Max Planck Institute for Psycholinguistics in The Netherlands. Two cameras filmed the front of each participant, while two more cameras filmed each of their bodies from a 45-degree angle, and two more cameras recorded a birds-eye view. Finally, one camera filmed the two participants together. See Figure 1 for an example image depicting the data recording set-up. All cameras filmed at 25 frames per second. Each participant's speech was recorded by a directional microphone. All video and audio files were synchronized, leaving a time resolution of 40 milliseconds (ms). Both video and audio files were utilized for the present study. Following the conversation

Figure 1. Still frame from the scene-view camera, illustrating the set-up of the conversational setting.

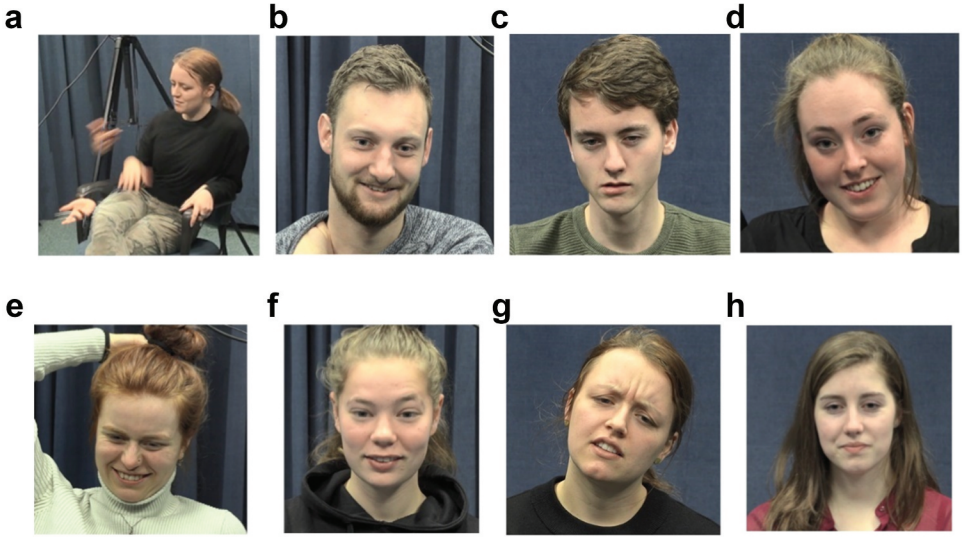

Figure 2. Examples of common visual signals. Panel **A** depicts a manual gesture as it unfolds; Panel **B**, smile; Panel **C**, squint; Panel **D**, eye-widening (with co-occurring eyebrow raise); Panel **E**, nose wrinkle (with co-occurring frown); Panel **F**, eyebrow raise; Panel **G**, frown; and Panel **H**, mouth corners down.

tasks, participants completed two questionnaires: the empathy quotient (EQ; Baron-Cohen & Wheelwright, 2004; Groen et al., 2015) and the brief version of the Fear of Negative Evaluation Scale (Leary, 1983; Watson & Friend, 1969). Questionnaire response data were not used for this study.

Informed consent was obtained before and after filming. The present study was conducted within existing ethical approval for corpus studies, approved by the Ethics Committee of the Social Sciences department of Radboud University, Nijmegen (approval code: ECSW 2018–124). Participants who appear in Figures 1 and 2 additionally gave written, informed consent to having identifiable images used for publication of this work.

Annotation of questions

Rather than analyzing all turn segments within the corpus, we focused on question–response pairs because questions are incredibly common in conversation and are part of so-called *adjacency pairs*. This context is particularly relevant for this current study, given that the interactional embedding of the utterance is crucial to our hypotheses. In other words, speaker A is producing a multimodal utterance *for* speaker B, who is expected to provide a relevant and timely response. This interaction is in contrast to utterances with much less of an expectation of response, such as self-directed questions, rhetorical questions, or quotations of questions asked by characters during narrative telling. Given that our hypotheses are framed around research on parallel planning and comprehension, this embedding in question–response adjacency pairs (Schegloff, 2007) is crucial. As described in Nota et al. (2021) and Trujillo and Holler (2021), questions and responses were manually annotated using the following procedure. First, we made an automatic orthographic transcription of the speech signal using the Bavarian Archive for Speech Signals Webservices (Kisler et al., 2017). Questions were identified and coded in ELAN (5.5; (Wittenburg et al., 2006)), largely following the coding scheme of Stivers and Enfield (Stivers & Enfield, 2010). In addition to this scheme, more rules were applied on an inductive basis in order to account for the complexity of the data in the corpus. Specifically, a holistic approach was adopted that took into consideration visual bodily signals, context, phrasing, intonation, and addressee behavior. Nonverbal sounds were excluded (e.g., laughter, sighs). This work was done by two human coders, one native speaker of Dutch, and one highly proficient speaker of Dutch. Interrater reliability between the two coders was calculated with raw agreement (Cohen, 1960; Landis & Koch, 1977) and a modified Cohen's kappa using EasyDIAg (Holle & Rein, 2015) on 12% of the total data (4 dyads, all tasks). A standard overlap criterion of 60% was used. Reliability between the coders resulted in a raw agreement of 75% and $k = 0.74$ for questions, and a raw agreement of 73% and $k = 0.73$ for responses, indicating substantial agreement. This approach resulted in a total of 6778 questions (duration Mdn = 1114 ms, range = 99–13,145 ms, IQR = 1138 ms). Because not all questions received a response, the present analyses are based on a subset of 4436 question–response sequences.

Annotation of visual signals

For the present analyses, visual signals were annotated in ELAN (5.5; Wittenburg et al., 2006) based on the synchronized frontal view videos from the corpus and linked to the question and response transcriptions. Only facial signals that started or ended between a time window of 200 ms before the onset of the question transcriptions and 200 ms after the offset of the question transcriptions were annotated. Manual annotations were created on a frame-by-frame basis by trained human coders.

For the hands data, we coded all co-speech, manual gestures, including representational and nonrepresentational gestures. In this case, we defined *representational co-speech gestures* as hand movements that visually depict or deictically refer to some aspect of what the speaker is talking about (Alibali et al., 2001). *Representational gestures* could be *iconic gestures* that represent an action (e.g., producing a hammering motion when talking about hammering) or object (e.g., using the hands to trace the outline of a circle, when referring to a ball); *pointing gestures* that refer to specific people or places referred to in speech; or *metaphoric gestures* that visually represent abstract concepts (e.g., holding the thumb and index finger close together to indicate smallness while uttering "I had so little time"). *Nonrepresentational gestures* included *interactive gestures*, such as an invitation for the addressee to respond (e.g., holding the palm flat and facing up, with fingers pointing to the addressee) (Bavelas et al., 1992), or *pragmatic gestures*, such as to distance one's self from what is said (e.g., raising the hands in front of one's chest, palms facing forwards) (Kendon, 2017). *Self-adaptors* (e.g., fixing one's hair) and *instrumental actions* (e.g., tying shoe laces) were not included. Other manual gestures, such as *beat gestures*, were not included. As with facial signals, *manual gestures* were coded starting with the first frame where hand or finger movement is evident, and ending when the hands or fingers

returned to their (new) rest position. *Gestures with multiple strokes* (i.e., the most meaning-carrying component, McNeill, 1992) produced in direct succession that carry the same meaning were coded as one single gesture. For example, if a person raises a fist and produces several up-and-down movements to depict hammering, this motion was counted as one gesture. *Gestures with consecutive strokes* with different meanings were split into multiple independent annotations. Boundaries between these gestures, when there was no clear return to rest position, was based on changes in handshape, motion, or placement of the hands. In total, 362 manual gestures occurring during questions were annotated. See Figure 2a for an example of a manual gesture.

For the face, we focused on the following signals: eyebrow frowns (n = 895), eyebrow raises (n = 1503), eyebrow–frown raises (n = 138), unilateral eyebrow raises (n = 204), eye widenings (n = 286), squints (n = 771), long blinks (i.e., longer than 410 ms (Hömke et al., 2018); n = 833), direct gaze (i.e., gaze toward the addressee, based on pupil position; n = 3029), nose wrinkles (n = 87), and non-articulatory mouth movements: pressed lips (n = 45), mouth corners down (n = 48), and smiles (n = 1562). See Figure 2b-h for examples of common visual signals that were coded. Movements due to swallowing, inhaling, laughter, or articulation were not considered.

Presence and prominence of visual signals

For our statistical analyses involving visual signals, we use two measures: the *number* of visual signals that occurred during an utterance half, and the *mean relative duration* of those signals (described in more detail under the *Statistical Analyses* subsection). These measures differ from the more information–theoretic measures used for the spoken language (described in following text) because currently no measures are comparable to semantic predictability of words. We therefore take the presence of visual signals as capturing increasing communicative information. Because there is also evidence that communicative relevance influences the duration of visual signals (Holler et al., 2022), we included mean relative duration to also capture the (temporal) prominence of these visual signals.

Similar to the questions and responses, interrater reliability for visual signals was calculated with raw agreement (Cohen, 1960; Landis & Koch, 1977) and a modified Cohen's kappa (*k*; Holle & Rein, 2015) using a standard overlap criterion of 60%. For manual gestures reliability was calculated based on 22.3% of the data (one randomly chosen segment of 5 minutes from each pair) that were coded by both coders (MtB, LvO). These randomly chosen segments contained 21.4% of the relevant representational gestures that were identified by the first coder (n = 60). For 37 of 68 participants, both coders agreed they did not gesture in the segment. For the other segments, annotations from the two coders were matched if they overlapped 60%, following Holle and Rein (2015). In other words, gestures were considered matching if both coders identified a gesture, and the second coder's annotation overlapped with the first coder's annotation by minimally 60% in terms of duration of the annotations (this approach is the ELAN default setting and standard procedure for gesture reliability, see (Holle & Rein, 2015). Because the first coder (MtB) annotated all gestures in this dataset, their annotations were used as the reference, and the second coder's (LvO's) annotations were compared with the first coder's annotations. For gesture identification, we observed 79.7% raw agreement. Cohen's kappa could not be calculated because there was only one gesture category. For reliability of facial signals, one question and one response in one of the three parts were selected randomly for each participant in all dyads. This approach enabled us to compare all coders in a pairwise fashion on the same data. For all facial signals, we calculated the range and average agreement and Kappa values for all pairs of coders. the paired comparisons showed an average raw agreement of 76% (min = 70%, max = 82%) and an average Kappa of 0.96 (min = .94, max = .97), indicating almost perfect agreement. See Nota et al. (2021) for more extensive methodological details on facial signal coding, reliability, and annotating timing precision. Note that somewhat different procedures for the two types of signals were used because the study design included multiple coders working on the facial signals.

Automatic transcription of questions

To calculate lexical information density, we first performed automatic speech recognition on our data, using an openly available Kaldi (Povey et al., 2011) implementation for Dutch (available: https://github.com/opensource-spraakherkenning-nl/asr_nl). Because automatic transcription programs are typically not without errors, two researchers (JPT, MK) manually checked a subset of the transcriptions to verify that the transcribed speech belonged to the current speaker (i.e., it was not due to overlapping speech from the other speaker), and that there were transcribed words for all instances of audible speech.

Trigram model of lexical information density

Before calculating our main measures of linguistic information, we first trained a second-order Markov (i.e., trigram) model, which formed the basis for our surprisal and entropy calculations (described in following text). Calculations were done using in-house developed Python (v.3.7 Van Rossum & Drake, 2009) scripts. We used the *nltk* python package (Bird et al., 2009) to train trigram (i.e., second-order Markov) models of word co-occurrence on the *Corpus Gesproken Nederlands* [Corpus of Spoken Dutch] (CGN; Van Eerten, 2007), using only the corpora involving dialogue (i.e., components A-D: face-to-face spontaneous conversation, interviews, telephone dialogs). This approach led to an inclusion of 5,277,612 words, and 609,508 utterances in our training corpus. Before training, non-lexical items (e.g., "uh," "oh") were removed, and transcribed contractions were converted to their full form (e.g., "d'ruit" – > "daaruit," "'m" – > "hem"). Trigram models take each utterance in the training corpus and count how many times a given word occurs after a given two-word sequence. For example, in English, the two-word sequence "Are you" could be followed by "going," "done," "ready," or other such terms. By counting the number of times each possible third word occurs after a given sequence relative to the number of possible third words, we can calculate the probability of that third word occurring. These probability distributions were then used for our *surprisal* and *entropy calculations.*

Surprisal as a measure of informativity

To obtain a measure of informativity at each point in an utterance, given the context of the previous two words, we can take the negative log-odds of the probability of a given third word occurring. See Eq.(1) for the definition of surprisal at word t given its context. To this end, we converted the trigram probability values to *surprisal* by taking the negative log of each probability, as calculated on the CGN corpus. Then, we extracted the corresponding surprisal value from each three-word sequence in the CoAct corpus.

$$S(w_t) = -\log P(w_t|w_1 \ldots_{t-1}) \tag{1}$$

Next-word entropy as a measure of lexical uncertainty

To obtain a measure of the uncertainty at each point in an utterance, given the context of the previous two words, we can calculate the Shannon entropy over the entire probability distribution of possible next-words (Aurnhammer & Frank, 2019; Van Schijndel & Linzen, 2019), providing us with a measure of *next-word entropy*. Therefore, after training our model on the CGN, we extracted the entropy value for each two-word sequence in the CoAct corpus. Eq. 2 provides a formal definition of next word entropy.

$$H(t) = \sum_{w_{t+1} \in W} P(w_{t+1}|w_1 \ldots_t) \log P(w_{t+1}|w_1 \ldots_t) \tag{2}$$

Data pre-processing

We excluded any utterances that were shorter than 600 ms in duration or contained fewer than eight words. This approach ensured that the utterances were sufficiently long such that parallel planning could conceivably occur during the speech and that the number of independent data-points was sufficient for the trigram calculations. The 600-ms threshold was chosen based on the finding that 600 ms is the minimum time required to plan a single word (Indefrey, 2011; Indefrey & Levelt, 2004). This step resulted in 886 utterances. Additionally, we excluded all utterances that had a mean transcription confidence less than 0.8 to ensure that only accurate transcriptions were included in the final analyses. This step resulted in 816 utterances. As a final pre-processing step, given our hypotheses regarding information distribution facilitating rapid turn-taking, we focused our analyses on questions with a clear response expectation. This focus was based on the social action that the question performs, with self-directed questions and active participation questions being excluded. After all pre-processing steps, we had 719 utterances for the final analyses. These utterances were accompanied by 1278 visual signals, consisting of 86 manual gestures and 1192 facial signals.

Statistical analyses

All statistical analyses were performed in R (R Core Team, 2019). Linear mixed models were implemented in lme4 (v.1.1–27.1; Bates et al., 2015), and *ggplot* (v.3.3.5; Wickham & Chang, 2014), *sjPlot* (v.2.8.4; Lüdecke, 2018), and *raincloudplots* (v.0.2.0; Allen et al., 2021) were used for visualization.

For our test of whether questions occurring in spontaneous conversation show higher surprisal or entropy in the first half or last half of an utterance, we used a linear mixed modeling approach. Specifically, we first split utterances into a first and last half based on duration, with surprisal and entropy based on the words in each utterance half. For words occurring at the split-point, the word was included in the first half if the split-point was closer to the onset of the word, and the word was included in the last half if the split-point was closer to the end of the word than the onset. We then built a null model that contained surprisal or entropy as the dependent variable, *utterance duration* and *number of words* as fixed effects, and dialogue (i.e., combination of dyad and participant) as a random intercept. Utterance duration and number of words were included as fixed effects given the potential variation in surprisal and entropy values that could arise due to differences inherent to the length of an utterance, and we chose to use both of these complementary variables given that turn duration and speech rate have been shown to be correlated with turn transition timing (Roberts et al., 2015). We then added utterance half (i.e., first or last; treatment coded, with first half as the reference level) and the main fixed effect of interest, and we used likelihood ratio test of model comparison to determine if adding utterance half led to a significant increase in the explained variance.

Our second test was whether the number of visual signals differed between the first and last half of an utterance. We used the same mixed effects modeling approach described previously, but with number of visual signals as the dependent variable. For this model, random intercepts were fitted for dyad/participant as a nested term.

Next, we tested how surprisal and entropy were associated with visual signaling by using separate models for the two (surprisal and entropy) dependent variables. This modelling was done by focusing on the utterance half that was found to be higher. In other words, if surprisal or entropy was higher in the first half of utterances, we tested how entropy in the first half of an utterance is associated with the number of visual signals. Specifically, we first tested whether surprisal or entropy of the first half was associated with the number of visual signals that occur in the first half of the utterance. We again constructed a linear mixed model, with *number of early signals* as the dependent variable, *utterance duration* and *number of words* as fixed effects, and *dyad/participant* as a nested random effect. We

again compared this null model to a model also containing *beginning-of-turn entropy* using a likelihood ratio test.

To determine if there are more fine-grained associations between visual signaling and next-word entropy, we tested whether entropy was associated with the presence of particular visual signals. Manual gestures were considered as one category, whereas each facial signals represented its own category. This approach was chosen because the manual gesture types are based on functional, qualitatively defined distinctions of what the gesture is contributing to the utterance, whereas the facial signal types are based on different muscle activations of specific facial articulators.

This determination was done by testing a mixed model with entropy as dependent variable, and *signal* as fixed effect. We again included *utterance duration* and *number of words* as additional fixed effects, and *dyad/participant* as a nested random effect. *Signal* was given as a categorical variable with the individual visual signals as factor levels, using sum coding. To test the model, we compared against the null model that did not contain *signal.*

Finally, given that *number of signals* does not consider the salience of these signals, we tested whether number of signals was associated with the mean relative duration of visual signals as an additional step. Mean relative duration was calculated by taking the duration of all signals that occurred during a particular beginning-of-turn and dividing their duration by the duration of the utterance, then taking the average of these relative durations. We then created a mixed model with *number of signals* as dependent variable and *mean relative duration*. This model allowed us to test whether the two measures of visual signaling (i.e., number and duration) show parallel or complementary patterns of use. *Parallel use* would be finding that signals are also longer when there are more signals, whereas *complementary use* would be finding that signals are on average shorter when there are more signals. We again used alikelihood ratio test to compare this model against the model without *mean relative duration*. For both models in this comparison, utterance duration and number of words were included as fixed effects, and dyad/participant was modeled as a nested random factor.

Post-hoc analyses

We additionally carried out several post-hoc analyses that were not part of our a priori hypotheses. These assessments consisted of a more fine-grained analyses of the social actions that the utterances were performing and compared entropy and visual signaling in the last half of the utterance; in addition, an association between first-half entropy and last-half visual signaling was analyzed. In the interest of space, we include these analyses in the Supplementary Materials.

Transparency of analytic methods

The quantitative data that support the findings of this study and primary analysis scripts are openly available in the Open Science Framework at https://osf.io/behuy/.

Not all participants agreed to their data being shared publicly, so raw video and transcript data are not available.

Results

Lexical information distribution within utterances

In line with our hypotheses, we found that surprisal significantly differed between the first and last half of the utterance, $\chi^2(1) = 89.409$, $p < .001$. Specifically, surprisal was found to be significantly lower in the last utterance half compared with the first (95% confidence interval [CI] for surprisal difference: [−0.737, −0.487], $t = -9.604$). See Figure 3a. Note that 70.9% of the utterances show a larger first-half surprisal compared with the last-half surprisal. Also in line with our hypotheses, we found that entropy significantly differed between the first and last utterance halves, $\chi^2(1) = 953.98$, $p < .001$. Specifically,

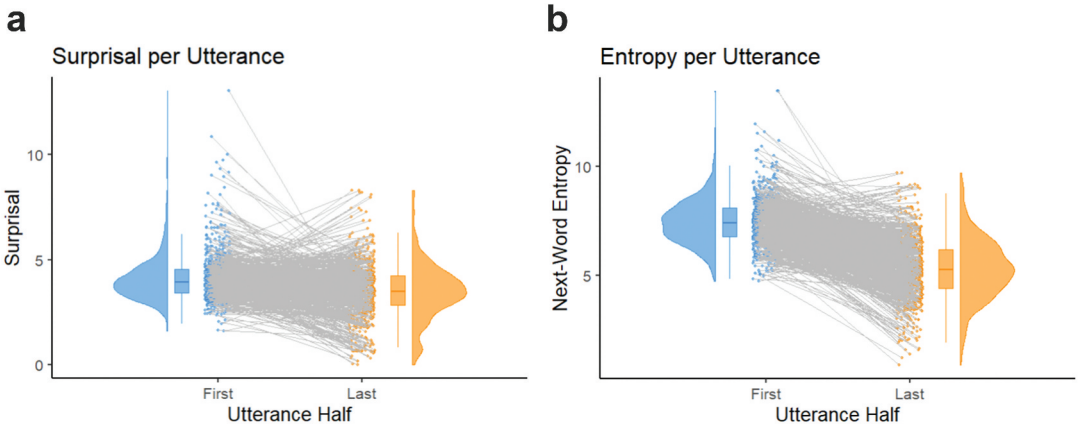

Figure 3. Surprisal and next-word entropy between utterance halves. Colored curves represent the smooth probability distribution for the data, and boxplots display the median (center line) and interquartile range (hinges). Whiskers on the boxplots extend to the furthest data point that is maximally 1.5 times the interquartile range away from the hinge. Individual data points are depicted as colored circles, with lines connecting datapoints from the same utterance, between first and last half of the utterance. In Panel **A**, surprisal is given on the y-axis. In Panel **B**, next-word entropy is given on the y-axis. In both panels, utterance half is given along the x-axis.

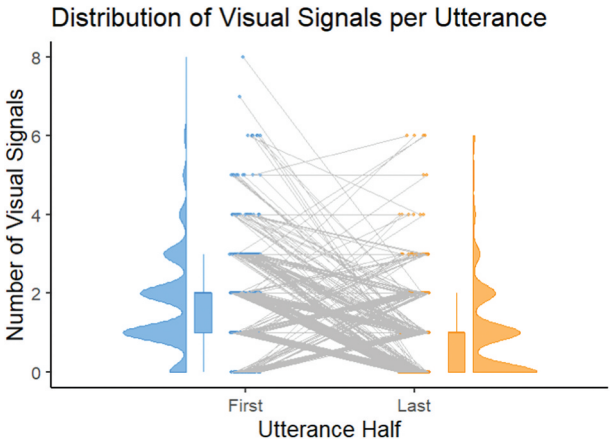

Figure 4. Number of visual signals (based on face and manual gestures) between utterance halves. Colored curves represent the smooth probability distribution for the data, and boxplots display the median (center line) and interquartile range (hinges). Whiskers on the boxplots extend to the furthest data point that is maximally 1.5 times the interquartile range away from the hinge. Individual data points are depicted as colored circles, with lines connecting datapoints from the same utterance, between first and last half of the utterance. Number of signals is given on the y-axis, while utterance half is given on the x-axis.

entropy was found to be significantly lower in the last utterance half compared with the first (95% CI for entropy difference: [−2.29, −2.06], t= −36.22). See Figure 3b. Note that 92.3% of the utterances show a larger first-half entropy compared with last-half entropy.

Visual signal distribution within utterances

We additionally found that the number of visual signals differed between the first and last utterance halves, $\chi^2(1) = 85.476$, $p < .001$. Specifically, more visual signals occurred in the first half compared with the last half (95% CI for difference in number of signals: [−0.591, −0.387], t= −9.363). Note that 64.6% of all utterances that contained at least one visual signal (based on facial signals and manual gestures) (56.5% of all utterances) followed this pattern. See Figure 4.

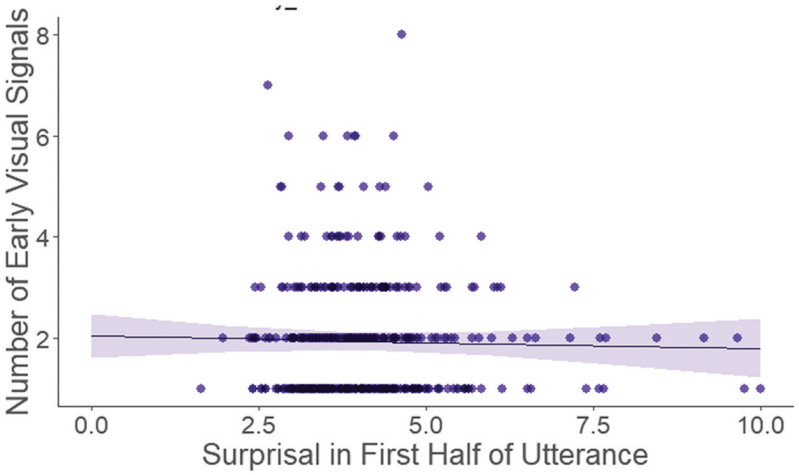

Figure 5. Association between number of visual signals and surprisal in the first half of an utterance. The y-axis provides number of signals, and the x-axis provides surprisal. Individual datapoints are depicted as gray circles. A fit line is provided, based on the marginal effects of our mixed effects model, with an 89% confidence interval indicated with gray shading.

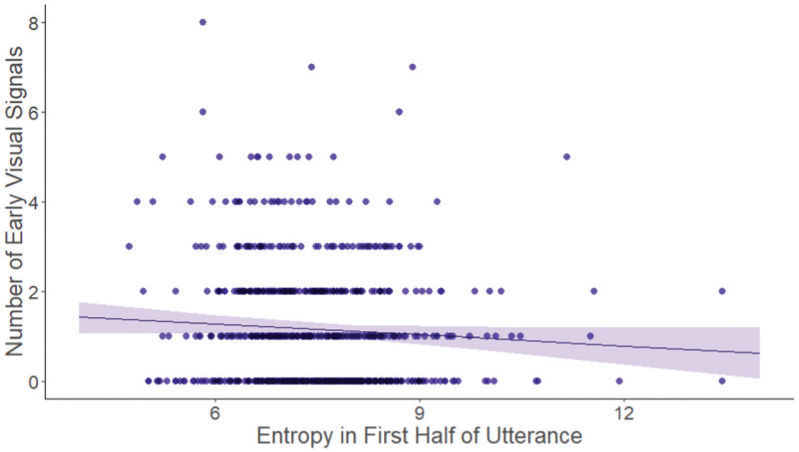

Figure 6. Association between number of visual signals and next-word entropy in the first half of an utterance. The y-axis provides number of signals, and the x-axis provides next-word entropy. Individual datapoints are depicted as gray circles. A fit line is provided, based on the marginal effects of our mixed effects model, with an 89% confidence interval indicated with gray shading.

Association between linguistic information and visual signaling within first utterance half

We tested whether surprisal in the first half of the utterance was associated with the number of visual signals that occur during the same time window. However, we found no association between surprisal and number of visual signals, $\chi^2(1) = 0.038$, $p = .845$. See Figure 5. We next tested whether entropy in the first half of the utterance was associated with the number of visual signals that occur during the same time window. We found a significant association between first-half entropy and first-half visual signals, $\chi^2(1) = 4.312$, $p = .038$. Specifically, higher entropy in the first half of the turn was associated with fewer visual signals (95% CI for reduction in number of visual signals per unit increase in entropy: [−0.109, −0.003], $t = -2.077$). See Figure 6. Finally, we tested whether surprisal or entropy in the first half of the utterance was associated with the presence of any particular signal. However, we found no evidence for such an association in surprisal, $\chi^2(17) = 21.155$, $p = .219$, or entropy, $\chi^2(17) = 17.791$, $p = .402$.

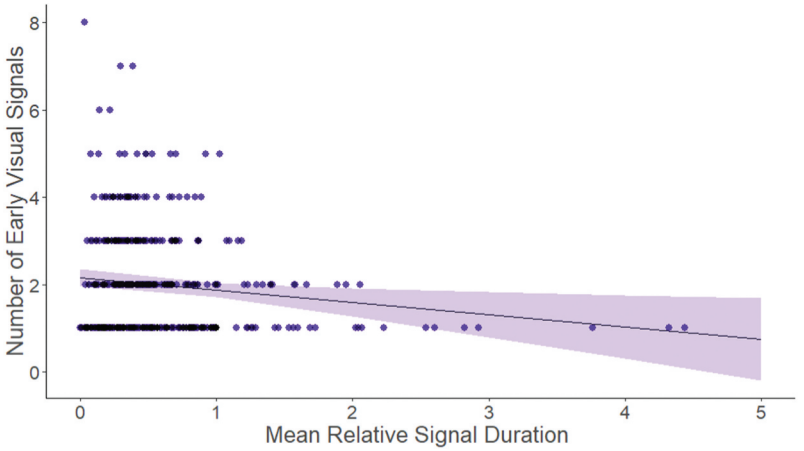

Figure 7. Association between number of visual signals and mean relative duration of those signals in the first half of an utterance. The y-axis provides number of visual signals, and the x-axis provides mean relative duration (i.e., mean of signal durations minus utterance durations). Individual datapoints are depicted as purple circles. A fit line is provided, based on the marginal effects of our mixed effects model, with an 89% confidence interval indicated with purple shading. Note that mean relative duration is often greater than 1 (note: 1 = utterance duration) as many signals start before speech onset and/or continue after the utterance has completed (see Nota et al., 2021).

Signal prominence: associations between number and duration of visual signals within first utterance half

Next, we tested whether there is an association between the number of visual signals in the first half of the utterance and the mean relative duration of those signals, while still taking into account the association between entropy and number of visual signals. In this analysis, we found a significant improvement in model fit when including mean relative duration, $\chi^2(1) = 6.890$, $p = .009$. Specifically, a lower number of visual signals was associated with a higher mean relative duration of those signals (95% CI for reduction in number of signals per mean second of duration increase: [−0.495, −0.072], $t = -2.624$). See Figure 7.

Discussion

This study aimed to determine how lexical information density is distributed across utterances spoken during natural face-to-face conversation between Dutch speakers, and how predictability pairs with the use of visual signals, such as facial signals and manual gestures. In terms of predictability, we found that both surprisal and entropy are higher in the first half of utterances (i.e., lower predictability than in the second half). We additionally found that visual signals occur more frequently in the first half of utterances. However, *within* the first utterance half, higher entropy is correlated with fewer, but relatively longer, visual signals. We found no evidence for an association between linguistic predictability and visual signaling in the last half of the utterance (see Supplementary Materials).

Our finding of higher surprisal and entropy density in the first utterance half is in line with previous work showing a gradient, or non-homogenous, distribution of information at the sentence or utterance level. Klafka and Yurovsky (2021) showed that across many languages, there is a smooth but often either increasing or decreasing distribution of information (quantified using surprisal) in written sentences. In other words, there are no sharp peaks and troughs in the distribution, even when there is a general trend of increasing or decreasing over time. The same study also showed increasing surprisal across spoken utterances in English (Klafka & Yurovsky, 2021). Similarly, Trujillo and Holler (2024b) investigated spoken

dialogue and showed that, across six languages, information distribution (again using surprisal) was either front-loaded (i.e., higher in the first half) or back-loaded (i.e., higher in the last half).

The current study further builds on this work by demonstrating that Dutch falls into the front-loaded language group, similar to German, Japanese, and Arabic (Trujillo & Holler, 2024b). This resutl corroborates the text-based findings from Klafka and Yurovsky (2021), who suggested that more phylogenetically related languages have more similar information distributions. Following from the results of Trujillo and Holler (2024b), it may be that spoken Dutch, similar to German, Arabic, and Japanese, also has an uneven distribution of nouns and verbs between the first and last half of utterances. In this case, it may be interesting for future research to quantify the occurrence of both visual signals of various kinds, as well as different grammatical units, to gain a more complete understanding of how these different components pattern onto spoken utterances, and whether cross-linguistic differences exist in the multimodal distribution of information. Finally, our results further show that this pattern of front-loaded information holds for both surprisal (i.e., informativity) and entropy (i.e., uncertainty), and in *spoken face-to-face* dialogue. These findings call into question the "given-before-new" principle, a long-claimed universal principle of information structure stating that speakers begin their utterances with old, or given, information, and new information *follows* the given information (Halliday, 1967). At least at the utterance level, in Dutch spontaneous face-to-face conversation, this pattern does not seem to hold.

We additionally found that visual signals of the face and hands occur more frequently in the first half of an utterance. This finding is in line with hypotheses posed in the theoretical framework by Holler and Levinson (2019), and empirical findings by Nota et al. (2021), who similarly reported that facial signals seem to occur most frequently near the beginning of a spoken utterance. The present study builds on this work by providing a statistical test showing that this pattern holds when including both manual gestures as well as facial signals. Importantly, the finding of higher visual signaling in the first half of utterances parallels our finding of higher surprisal and entropy in the first utterance half. This result suggests that the first half of an utterance contains the most visual information and the most informative and unpredictable lexical content. Together, these results suggest strong front-loading of multimodal information during dialogue in Dutch speakers and align with the idea that such multimodal front-loading may benefit early next turn planning during conversation (Holler & Levinson, 2019).

Although the general pattern in our findings was of multimodal front-loading of information, we also observed a more dynamic balance between visual and lexical information. Specifically, we found that the use of more, but relatively shorter, visual signals was associated with lower first half-of-turn entropy. This negative association between number and relative duration of signals suggests that when quantifying the prominence or degree of visual signaling, we must consider both number and duration of signals. In this case, it is possible that this balance between number and duration is a way to reduce the amount of information (i.e., by not employing too many different visual components) while simultaneously ensuring that those fewer visual components are being picked up by the addressee, despite—or perhaps precisely due to—the high uncertainty at the level of lexical items. While previous work (Nota et al., 2021), as well as the global analysis of the present study, show that there are generally more visual facial signals early in an utterance versus later, our results expand on this research by showing that the number of signals occurring in a given utterance may systematically relate to the overall distribution of information within an utterance, such that higher information density in the verbal channel tends to be associated with a lower number, but longer lasting, visual signals from the face and hands.

Our results therefore suggest a complementarity in the distribution of spoken and visual signals. In other words, speakers may aim to minimize the effort of production and/or comprehension by balancing high lexical uncertainty with temporally prominent visual signals that can support comprehension (Zhang et al., 2021), while minimizing the number of different visual signals, which could otherwise increase comprehension effort due to the many information sources.

An important nuance to our findings of high information density relating to a lower number of—but more prominent—visual signals, is that there was only evidence for such an association in the first half of the utterance. In the last half of the utterance, we found no association between surprisal or entropy and number of visual signals (see Supplementary Materials). The number of visual signals in the last half of the utterance also did not correlate with entropy in the first half of the utterances. Although these analyses were post-hoc in nature and thus should be interpreted with some caution, they suggest that there is a tighter coordination between the visual and linguistic channels during the portion of the utterance with less predictable linguistic information. These results may be relevant for models of conversational turn-taking, whereas previous results suggest that next-speakers begin to plan their response in parallel with listening to an ongoing utterance, as soon as sufficient information has been provided (Barthel & Levinson, 2020; Barthel et al., 2017; Bögels, 2020; Bögels et al., 2015). In spoken Dutch, this early response planning may be facilitated by the high multimodal-information density observed in the present study. The way visual signals are employed, in terms of their number and duration, may also be important for the listener's parallel planning and listening. Experimental work is needed to test how number and prominence of visual signals, together with linguistic uncertainty, jointly affect language comprehension.

That we did not find an association between mean surprisal and visual signaling suggests that, in the first half of an utterance, speakers are not adding additional layers of information in already information-dense stretch of speech, as would be the case if first half-of-turn surprisal positively associated with number of visual signals. In contrast, visual signal usage (in the first half of the utterance) is modulated by how unpredictable (i.e., as measured by entropy) the speech is (Zhang et al., 2021), with few, but prominent signals helping to reduce this uncertainty. This subtle distinction is particularly interesting for demonstrating the complementary insights from surprisal and next-word entropy as measures of linguistic predictability.

The lack of association between surprisal and visual signals may also be related to the level of analysis. For example, the association between visual signals and surprisal has also recently been explored by Grzyb et al. (2022). In that study, high surprisal words were more likely to be accompanied by an iconic gesture. This association between high surprisal words and iconic gesture use therefore seems at odds with the current findings. However, there are several important methodological differences between these two studies that may provide further insights into how the different outcomes can be accounted for. First, Grzyb and colleagues looked at word-level predictability and iconic manual gestures, whereas we investigated mean predictability over a half of the utterance and considered all representational manual gestures, as well as facial signals. Another important difference is that the surprisal-iconic gesture association was found in English corpora (Grzyb et al., 2022), and thus cross-linguistic differences in information distributions (Trujillo & Holler, 2024b) could also extend to differences in multimodal information distribution. Future work should aim to determine to what extent the multimodal patterns observed in our study can be generalized across languages, and how utterance-level distribution patterns relate to moment-to-moment (e.g., word-level) associations.

Limitations and future directions

One important limitation to the current work is that our method of calculating surprisal and next-word entropy led to our analyses focusing on a subset of all spoken utterances, whereas very short utterances were excluded. It is, of course, entirely possible that shorter utterances show a different pattern of information distribution or that they pattern differently in terms of multimodal information. To address such possibilities, complementary methods of calculating lexical information will be needed, utilizing a larger range of utterance lengths. Additionally, while we were able to capture a range of visual signals that included both facial signals and manual gestures, other signals, such as head gestures or body posture, will likely also contribute to the total multimodal array of utterance information. Finally, the current study did not consider speech prosody, which also plays an important role in information structure, and will likely also contribute to informativeness and surprisal.

Similarly, co-articulation (i.e., visible speech) provides early cues to upcoming words, thus shaping the actual surprisal and next-word entropy during the comprehension of natural speech. The present study therefore provides a starting point for investigating how information is distributed multimodally, but should be built upon by studies that take into account the information that is also available in the acoustics of speech, as well as the informativity of other visual signals. Another particularly interesting avenue for future research will be to determine if this association between visual signaling and lexical entropy holds in languages with different information distribution patterns, such as English, Spanish, or Mandarin (Trujillo & Holler, 2024b).

Finally, despite appearing challenging at this time, future research should seek to find methods of quantifying the informativity of visual signals in a way that is comparable, at least for analytical purposes, to lexical or semantic predictability of words. This approach would allow a more precise assessment of how visual and lexical (or semantic) information is distributed within an utterance.

Conclusions

The current study thus shows that in spoken Dutch face-to-face conversation, visual signals as well as less predictable information in the spoken modality are more densely distributed into the first half of an utterance. Utterances with high uncertainty in the first half specifically tend to also be accompanied by fewer visual signals during this early portion of the utterance. When fewer visual signals are used, these signals also tend to be more visually prominent (in terms of relative duration). These findings have implications for bringing multimodality into the study of information distribution, providing a more complete view of communicative information and how multimodal information is coordinated, and they fit well with frameworks suggesting that the formation of multimodal utterances favors ease of utterance processing during conversational turn-taking.

Acknowledgments

We are grateful to Anne-Fleur van Drunen, Guido Rennhack, Hanne van Uden, Josje de Valk, Leah van Oorschot, Lina van Otterdijk, Maarten van den Heuvel, Mareike Geiger, Marieke Elderman, Marlijn ter Bekke, Michelle Kühn, Pim Klaassen, Rob Evertse, Veerle Kruitbosch and Wieke Harmsen for contributing to data collection, transcription checking, or annotating the visual signals.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by an ERC Consolidator grant [#773079, awarded to J. Holler].

ORCID

James P. Trujillo http://orcid.org/0000-0003-4713-376X

References

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, *44*(2), 169–188. https://doi.org/10.1006/jmla.2000.2752

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*, 63. https://doi.org/10.12688/wellcomeopenres.15191.2

Atkinson, J. M., Heritage, J., & Oatley, K. (1984). *Structures of social action*. Cambridge University Press.
Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198
Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175. https://doi.org/10.1023/B:JADD.0000022607.19833.00
Barthel, M. (2021). Speech planning interferes with language comprehension: Evidence from semantic illusions in question-response sequences. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*. SemDial25, Potsdam, Germany. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/11009
Barthel, M., & Levinson, S. C. (2020). Next speakers plan word forms in overlap with the incoming turn: Evidence from gaze-contingent switch task performance. *Language, Cognition and Neuroscience*, *35*(9), 1183–1202. https://doi.org/10.1080/23273798.2020.1716030
Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final "Go-Signals." *Frontiers in Psychology*, *8*, 8. https://doi.org/10.3389/fpsyg.2017.00393
Barthel, M., & Sauppe, S. (2019). Speech planning at turn transitions in dialog is associated with increased processing load. *Cognitive Science*, *43*(7), e12768. https://doi.org/10.1111/cogs.12768
Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.
Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, *15*(4), 469–489. https://doi.org/10.1080/01638539209544823
Bavelas, J. B., Gerwing, J., & Healing, S. (2014). Including facial gestures in gesture-speech ensembles. In M. Seyfeddinipur & M. Gullberg (Eds.), *From gesture in conversation to visible action as utterance: Essays in honour of Adam Kendon* (pp. 15–34). John Benjamins Publishing Company.
Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition*, *203*, 104347. https://doi.org/10.1016/j.cognition.2020.104347
Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia*, *109*, 295–310. https://doi.org/10.1016/j.neuropsychologia.2017.12.028
Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, *5*(1), Article 1. https://doi.org/10.1038/srep12881
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104
Cooperrider, K., Abner, N., & Goldin-Meadow, S. (2018). The palm-up puzzle: Meanings and origins of a widespread form in gesture and sign. *Frontiers in Communication*, 3. https://doi.org/10.3389/fcomm.2018.00023
Corps, R. E., Crossley, A., Gambi, C., & Pickering, M. J. (2018). Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition*, *175*, 77–95. https://doi.org/10.1016/j.cognition.2018.01.015
Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. *Discourse Processes*, *55*(2), 230–240. https://doi.org/10.1080/0163853X.2017.1330031
Domaneschi, F., Passarelli, M., & Chiorri, C. (2017). Facial expressions and speech acts: Experimental evidences on the role of the upper face as an illocutionary force indicating device in language comprehension. *Cognitive Processing*, *18*(3), 285–306. https://doi.org/10.1007/s10339-017-0809-6
Drijvers, L., & Holler, J. (2023). The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, *30*(2), 792–801. https://doi.org/10.3758/s13423-022-02178-x
Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384–392.
Ekman, P. (2004). Emotional and conversational nonverbal signals. In J. M. Larrazabal & L. A. P. Miranda (Eds.), *Language, knowledge, and representation* (pp. 39–50). Springer Netherlands.
Gerwing, J., & Allison, M. (2009). The relationship between verbal and gestural contributions in conversation: A comparison of three methods. *Gesture*, *9*(3), 312–336. https://doi.org/10.1075/gest.9.3.03ger
Groen, Y., Fuermaier, A. B. M., Den Heijer, A. E., Tucha, O., & Althaus, M. (2015). The empathy and systemizing quotient: The psychometric properties of the dutch version and a review of the cross-cultural stability. *Journal of Autism and Developmental Disorders*, *45*(9), 2848–2864. https://doi.org/10.1007/s10803-015-2448-z
Grzyb, B., Frank, S., & Vigliocco, G. (2022). Communicative efficiency in multimodal language. *PsyArXiv*. https://doi.org/10.31234/osf.io/a9wt3
Halliday, M. A. K. (1967). Notes on transitivity and theme in English: Part 2. *Journal of Linguistics*, *3*(2), 199–244. https://doi.org/10.1017/S0022226700016613
Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, *47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7

Holler, J., Bavelas, J., Woods, J., Geiger, M., & Simons, L. (2022). Given-new effects on the duration of gestures and of words in face-to-face dialogue. *Discourse Processes*, *59*(8), 619–645. https://doi.org/10.1080/0163853X.2022.2107859
Holler, J., & Beattie, G. (2002). A micro-analytic investigation of how iconic gestures and speech represent core semantic features in talk. *Semiotica*, *2002*(142). https://doi.org/10.1515/semi.2002.077
Holler, J., & Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica*, *2003*(146). https://doi.org/10.1515/semi.2003.083
Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review*, *25*(5), 1900–1908. https://doi.org/10.3758/s13423-017-1363-z
Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006
Holler, J., Tutton, M., & Wilkin, K. (2011, September). Co-speech gestures in the process of meaning coordination. *Proceedings of the 2nd GESPIN - Gesture & Speech in Interaction (Gespin)*, Bielefeld, Germany. https://research.manchester.ac.uk/en/publications/co-speech-gestures-in-the-process-of-meaning-coordination
Hömke, P., Holler, J., & Levinson, S. C. (2018). Eye blinks are perceived as communicative signals in human face-to-face interaction. *PLOS ONE*, *13*(12), e0208030.
Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297–315. https://doi.org/10.1037/a0022128
Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00255
Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144.
Jack, R. E., & Schyns, P. G. (2015). The human face as a dynamic tool for social communication. *Current Biology*, *25*(14), R621–R634. https://doi.org/10.1016/j.cub.2015.05.052
Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, *40*(4), 577–592. https://doi.org/10.1006/jmla.1999.2634
Kendon, A. (1985). Some uses of gesture. In D. Tannen & M. Saville-Troike (Eds.), *Perspectives on silence* (pp. 215–234). Ablex Publishing Corporation.
Kendon, A. (2017). Pragmatic functions of gestures: Some observations on the history of their study and their nature. *Gesture*, *16*(2), 157–175. https://doi.org/10.1075/gest.16.2.01ken
Kendrick, K. H., Brown, P., Dingemanse, M., Floyd, S., Gipper, S., Hayano, K., Hoey, E., Hoymann, G., Manrique, E., Rossi, G., & Levinson, S. C. (2020). Sequence organization: A universal infrastructure for social action. *Journal of Pragmatics*, *168*, 119–138. https://doi.org/10.1016/j.pragma.2020.06.009
Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*, *52*(4), 255–289. https://doi.org/10.1080/0163853X.2014.955997
Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005
Klafka, J., & Yurovsky, D. (2021). Characterizing the typical information curves of diverse languages. *Entropy*, *23*(10), Article10. https://doi.org/10.3390/e23101300
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310
Leary, M. R. (1983). A brief version of the fear of negative evaluation scale. *Personality and Social Psychology Bulletin*, *9*(3), 371–375. https://doi.org/10.1177/0146167283093007
Levinson, S. C. (2016). Turn-taking in human communication – Origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010
Levinson, S. C. (2017). *Speech acts*. The Oxford Handbook of Pragmatics. https://doi.org/10.1093/oxfordhb/9780199697960.013.22
Levshina, N., Namboodiripad, S., Allassonnière-Tang, M., Kramer, M., Talamo, L., Verkerk, A., Wilmoth, S., Rodriguez, G. G., Gupton, T. M., Kidd, E., Liu, Z., Naccarato, C., Nordlinger, R., Panova, A., & Stoynova, N. (2023). Why we need a gradient approach to word order. *Linguistics*, *61*(4), 825–883. https://doi.org/10.1515/ling-2021-0098
Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, *42*(S4), 1166–1183. https://doi.org/10.1111/cogs.12597
Lüdecke, D. (2018). *sjPlot—Data visualization for statistics in social science* [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.1308157
Magyari, L., Bastiaansen, M. C. M., de Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience*, *26*(11), 2530–2539. https://doi.org/10.1162/jocn_a_00673
Magyari, L., De Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00211
McHoul, A. (2008). Review of sequence organization in interaction: A primer in conversation analysis. Vol. 1 [Review of *review of sequence organization in interaction: A primer in conversation analysis. Vol. 1*, by E. A. SCHEGLOFF]. *Discourse Studies*, *10*(4), 576–581. https://doi.org/10.1177/14614456080100040602 JSTOR

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
McNeill, D. (2000). *Language and Gesture*. Cambridge University Press.
McNeill, D. (2016). *Why we gesture: The surprising role of hand movements in communication*. Cambridge University Press.
Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies*, *9*(2), 194–225. https://doi.org/10.1177/1461445607075346
Müller, C. (2004). The palm-up-open-hand. A case of a gesture family? In C. Müller & R. Posner (Eds.), *The semantics and pragmatics of everyday gestures* (pp. 233–256). Weidler. https://www.kuwi.europa-uni.de/de/lehrstuhl/sw/sw0/publikationen/Artikel-Mueller/Mueller_-FormsUsesPUOHand_2004.pdf
Nölle, J., Chen, C., Hensel, L. B., Garrod, O. G. B., Schyns, P. G., & Jack, R. E. (2021). Facial expressions of emotion include iconic signals of rejection and acceptance. *Journal of Vision*, *21*(9), 2932.
Nota, N., Trujillo, J., & Holler, J. (2022). *Conversational eyebrow frowns facilitate question identification: An online VR study*. PsyArXiv. https://doi.org/10.31234/osf.io/fcj8b
Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, *11*(8), Article 8. https://doi.org/10.3390/brainsci11081017
Nota, N., Trujillo, J. P., & Holler, J. (2023). Specific facial signals associate with categories of social actions conveyed through questions. *PLOS ONE*, *18*(7), e0288104. https://doi.org/10.31234/osf.io/qrhdf
Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, *19*(4), 605–616. https://doi.org/10.1162/jocn.2007.19.4.605
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (Eds.). (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.
Raymond, C. W. (2016). Sequence organization. In Jon Nussbaum (Ed.) *Oxford research encyclopedia of communication* (pp. 1–37). Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.133
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* (509th ed., Vol. 5). Frontiers in psychology. https://doi.org/10.3389/978-2-88919-825-2
R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/
Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I*. Cambridge University Press.
Snoek, L., Jack, R. E., Schyns, P. G., Garrod, O. G. B., Mittenbühler, M., Chen, C., Oosterwijk, S., & Scholte, H. S. (2023). Testing, explaining, and exploring models of facial expressions of emotions. *Science Advances*, *9*(6), eabq8421. https://doi.org/10.1126/sciadv.abq8421
Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, *39*(3), 392–417. https://doi.org/10.1006/jmla.1998.2586
Stivers, T. (2013). Sequence organization. In J. Sidnell & T. Stivers (Eds.) *The handbook of conversation analysis* (pp. 191–209). Blackwell Publishing Ltd.
Stivers, T., & Enfield, N. J. (2010). A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, *42*(10), 2620–2626. https://doi.org/10.1016/j.pragma.2010.04.002
Trujillo, J. P., & Holler, J. (2021). The kinematics of social action: Visual signals provide cues for what interlocutors do in conversation. *Brain Sciences*, *11*(8), 996. https://doi.org/10.3390/brainsci11080996
Trujillo, J. P., & Holler, J. (2023). Interactionally embedded gestalt principles of multimodal human communication. *Perspectives on Psychological Science*, *18*(5), 1136–1159. https://doi.org/10.1177/17456916221141422
Trujillo, J. P., & Holler, J. (2024a). Conversational facial signals combine into compositional meanings that change the interpretation of speaker intentions. *Scientific Reports*, *14*(1), 2286. https://doi.org/10.1038/s41598-024-52589-0
Trujillo, J. P., & Holler, J. (2024b). Information distribution patterns in natural dialogue differ across languages. *Psychonomic Bulletin & Review*, *31*(4), 1723–1734. https://doi.org/10.3758/s13423-024-02452-0
van Eerten, L. (2007). Over het corpus gesproken Nederlands. *Nederlandse Taalkunde*, *12*(3), 194–215. https://tsg-humlab.pages.science.ru.nl/cgn/
Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.
van Schijndel, M., & Linzen, T. (2019). Can entropy explain successor surprisal effects in reading? *Proceedings of the Society for Computation in Linguistics (Scil)*, *2019*, 1–7. https://doi.org/10.7275/qtbb-9d05
Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, *33*(4), 448.
Wickham, H., & Chang, W. (2014). *Ggplot2* [Computer software] Springer-Verlag New York. https://github.com/tidyverse/ggplot2
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN : A professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 1556–1559. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_60436
Yamashita, H., & Chang, F. (2001). "Long before short" preference in the production of a head-final language. *Cognition*, *81*(2), B45–B55. https://doi.org/10.1016/S0010-0277(01)00121-4

Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1955), 20210500. https://doi.org/10.1098/rspb.2021.0500

Appendix I. Instructions and topic prompts used for the corpus

(1) *Hoeveel privacy heb je nog tegenwoordig? [How much privacy do you have nowadays?]*
Tegenwoordig is je telefoon niet meer weg te denken uit het dagelijks leven. Je gebruikt hem onder andere om snel even wat op te zoeken, makkelijk te communiceren met vrienden en als navigatiesysteem. Maar alle apps die je gebruikt verzamelen ook heel veel data over jou. Dit wordt meestal voor positieve doeleinden gebruikt, zoals het opsporen van vermiste mensen of terrorismepreventie. Maar je moet niet vergeten dat systemen de indeling van jouw leven op deze manier heel goed in kaart kunnen brengen. Door locatiegegevens weten ze waar je bent en door zoekgegevens wat je leuk vindt. Verder verdienen bedrijven veel geld door jouw data te verkopen. Hoe sta jij tegenover het verzamelen van persoonlijke informatie? Welke data mag er wel en niet over jou verzameld worden? Hoeveel privacy ben jij bereid om op te offeren ten behoeve van gemak?

[These days you can't imagine not having your phone in daily life. You use it to look things up, easily communicate with friends, and as a GPS. But all the apps that you use collect data about you. This is often put toward good purposes, like tracking down missing people or preventing terrorism. But don't forget that these systems can easily use this to piece together how you live your life. Location information let them know where you are and search terms tell them what you like. Beyond this, companies earn a lot of money by selling your data. How do you stand regarding the collection of personal data? What personal data can and cannot be collected from you? How much privacy are you willing to sacrifice for the sake of convenience?]

(2) *Voor- en nadelen van social media [Pros and cons of social media]*

Facebook, Instagram, Twitter en WhatsApp . . . Social media is overal! Het is handig om snel met je vrienden te kunnen communiceren en om contact te houden met mensen die ver weg wonen en je niet zo vaak meer ziet. Helaas heeft social media ook nadelen, denk bijvoorbeeld aan cyberpesten of het verspreiden van nepnieuws of extremistische ideeën. Verder is er ook een grote kans op social-mediaverslaving. Wat vind jij voor- en nadelen van social media? Overweeg jij weleens je social media accounts te verwijderen? Vind je dat mensen te veel tijd online besteden in plaats van in normale sociale interactie, of maakt dit voor jou niet uit?

[Facebook, Instagram, Twitter, and WhatsApp. Social media is everywhere! It's useful for quickly communicating with friends and keeping in touch with people who live far away and you don't often see anymore. Unfortunately, social media also has downsides. For example, consider cyber-bullying or spreading fake news or extremist ideas. There's also a big chance of social media addiction. What do you think are pros and cons of social media? Do you ever consider deleting your social media accounts? Do you think people spend too much time online instead of in normal social interaction, or does it not matter to you?]

(3) *Studeren in het Engels of het Nederlands? [Studying in English or in Dutch?]*

Steeds meer universitaire opleidingen worden alleen nog in het Engels aangeboden. Engels is tenslotte de taal van de wetenschap en van het (internationale) bedrijfsleven. Bovendien trekt dit buitenlandse studenten aan, wat zorgt voor internationalisering van de campus. Maar de kwaliteit van het onderwijs gaat er niet per se op vooruit; lang niet alle docenten en studenten kunnen zich net zo goed uitdrukken in het Engels als in het Nederlands. En voor sommige banen is het júist belangrijk dat je je goed kunt uitdrukken in woord en geschrift in het Nederlands. Zou jij liever les krijgen in het Engels of in het Nederlands? En ben je bang dat de kwaliteit van het onderwijs hierdoor achteruit gaat? Of denk je dat een Engelstalige opleiding je carrièrekansen juist vergroten?

[More and more university educational programs are only being offered in English. English is, after all, the language of science and of the (international) industry. English courses also attract foreign students, which ensures internationalization of the campus. But the quality of education doesn't necessarily improve. By far, not all teachers and students can express themselves as well in English as they can in Dutch. And for some jobs it's actually necessary that you're able to express yourself in spoken and written Dutch. Would you rather take courses in English or in Dutch? And are you worried that the quality of education will be negatively affected by this? Or do you think that receiving an education in English can improve your chances on the job market?]