Do eyes make words? Do words see them?
The grammar of multi-modal interaction

Martina Wiltschko
ICREA, UPF
Draft, May 2021

Abstract.
An analysis is proposed for the form-meaning pairing in a minimal summons-answer adjacency pair: a prolonged stare answered by *what*. It is argued that the stare constitutes a genuine initiation move which licenses the use of *what.* It is shown that this use of *what* is restricted to answer initiation moves and cannot be used to react to any kind of situation. It is further shown that the interpretation of *what* in this context does not derive from ellipsis of a full sentence (i.e., *What are. you looking at*?). It is shown that the interpretation of *what* and the stare can straightforwardly be explained if it is assumed that these forms associate with a preconfigured structure which adds meaning to isolated units of language. The special role of eye-gaze is discussed: unlike words, which cannot be used outside of language, *eye-gaze* always has a dual function: to observe and to interact.

Keywords: summons, eye-gaze, interactional language, syntax-pragmatics interface

1 Introduction

When we talk to each other, we do not just use words, we use our bodies as well. This includes facial expressions, eye-gaze, gestures, and posture. Embodied interaction has long been the object of investigation in conversation analysis and related frameworks (see Mondada 2019 for a recent overview) but has played little role in formal grammatical analyses (but see Esipova 2019, Schlenker 2018, and Ginzburg et al. 2020). Conversation analysis and its kin is concerned with human interactions in their social context including embodied action. Formal grammatical analysis on the other hand is concerned with the formal properties of language. For formal grammatical analyses, the unit of analysis is the sentence; and its ingredients are the words that express concepts as well as the formal grammatical features which serve to glue these words together into meaningful sentences. Hence, neither what happens in interaction plays a role in theorizing about grammar nor do forms that are outside the realm of the "logos", such as eye-gaze or gestures.

The goal of this paper is to bridge the gap between these two approaches. I wish to demonstrate that grammar plays a critical role in the regulation of multi-modal communication. Thus, I argue that on the one hand, that formal theories of grammar should not ignore embodied language in interaction, as it provides a novel empirical domain, which can tell us much about the workings of the language faculty. On the other hand, I argue that to reach a comprehensive understanding of embodied interaction, conversation analysis benefits from the addition of formal grammatical analyses.

To show that multi-modal communication in interaction is regulated by grammar, I analyse a single adjacency pair, drawn from an existing corpus. Roughly, the adjacency pair I analyse, can be represented as in (1). The initiator (I) stares at their interlocutor. And this stare prompts the responder (R) to respond with a single-word utterance.

(1) I: 
R: What?

I show that this conversation, despite its apparent simplicity, displays properties that reveal the workings of an underlying system which enriches the meaning of the signs used. This system, I propose, is the grammar of interactional language (henceforth i-language) in the sense of Author (to appear a). I show that the eye-gaze serves to signal an initiation move and hence is to be analysed as a unit of language in its own right. Evidence for this analysis comes from the fact that the response marker *what* is only licensed in response to an explicit initiation move. It serves to question the nature of the initiation. Crucially, I show that this interpretation does not come about via eliding part of a full sentence pronouncing only *what* (i.e., *what* ~~*is your initiation*~~). Instead I argue that the full meaning derives from the grammar of interaction itself.

The paper is organized as follows. In section 2, I introduce the adjacency pair in (1) as it occurs in an actual conversation. In section 3, I introduce the framework and in section 4, I introduce the proposal. Sections 5 and 6 are dedicated to exploring the properties of *what* and the stare, respectively. In section 7, I conclude.

2 The problem: the form and function of summons-answer pairs

Consider the following interaction. The context is a family dinner with the mother (J), a sister (E) and her brother (T). During a conversation between J and T about T's bangs, E looks up and stares at T (01-02). Once T notices the stare, he responds with *what*. This sequence is the empirical focus in this paper.

(2)
01 E ((looks up at T))
02 (3.2)
03 T→ What.
04 (2.2)
05 E Nothing. I didn't
06 say anything.
07 T °Don't stare at°
08 me then
09 J St[op being so=
10 E [What?
11 J =aggressive Emily
12 E Actually Tom didn't
13 actually need to…

Clift: F1:6:24:19'

In terms of conversation analysis, the sequence of the stare followed by *what* is considered a *summons-answer* adjacency pair. Summons are moves used to attract the interlocutor's attention especially in situations when the initiator cannot be sure whether they have their attention (Schegloff 1968, 2007). They have also been referred to as 'attention-getting devices' (McTear,

1985) or 'attention-drawing devices' Keenan and Schieffelin 1976, Ochs et al. 1979). They include terms of address such as *Mommy*, certain courtesy phrases (*excuse me*), attention getting particles (*hey*) as well as certain questions (*you know what).* In addition to these verbal devices, gestures and eye-gaze may also be employed, either by itself or in addition to verbal clues (McTear, 1985, Ochs et al. 1979, Sacks 1992, Schegloff 1968, 2007). As for answers to summons, they typically involve verbal turns such as *yeah, what, uhuh* in combination with a redirection of gaze or posture (Jefferson 1972, Schegloff, 2007: 49). They function as *go-aheads* to encourage the interaction requested simply by means of giving the requested attention. This much is straightforward about the adjacency pair in (1): it functions as a typical summons-answer pair.

What I wish to address here is the question regarding the relation between the forms used and the function they receive in this context. Specifically, there are a number of non-trivial questions that arise. First, what is it about the stare that makes it a summons? The same question is raised for the answer: what exactly does *what* mean in the context of responding to a summons, and how is this meaning derived? Interestingly, E's response (in 05-06) denies that meaning is associated with her stare; she explicitly states that she didn't *say* anything. But T's next move contests this denial by telling E to stop the stare, suggesting that he does not agree with E's assessment. This is then followed by J's reprimanding E for being aggressive. Since there is nothing in the verbal interaction itself that would make it aggressive, the perception of aggression is likely to stem from the eye-gaze itself and perhaps other non-verbal clues. The entire sequence, then, raises a more general question: how do visual and verbal interaction combine to produce a coherent conversation? While conversation analysis provides the tools and insights to analyse the function and sequencing of conversational turns, it remains silent regarding the way these functions are mapped onto the forms that express them. This is the question I explore here.

3 The framework: the grammar of interactional language

The core proposal I develop is that the properties of the sequence in (1) are best understood as involving the working of an abstract system, a grammar of sorts. That conversations are regulated by a system which is part of our competence is the hallmark of conversation analysis. Author (to appear a), combines insights of conversation analysis with those of generative grammar. With an in-depth investigation of units of language (UoLs) that contribute to the interaction itself, rather than to its content, Author concludes that the same system which configures the content of interaction also configures the logic of the interaction itself. The UoLs under investigation are confirmationals (such as utterance-final *eh* and *huh*), which define initiating moves and response markers (such as utterance-initial *yeah* and *no*), which define reaction moves.

The core argument that there is a grammar of i-language stems from the fact that the class of confirmationals and the class of response markers display the same patterns of multi-functionality, and they do so across many unrelated languages. For example, the response markers can be used to answer questions, indicate agreement, merely acknowledge the move of an interlocutor or simply mark a response as a response. Author argues that this multi-functionality indicates the presence of an underlying abstract system (the so called i(nteractional)-spine, which enriches the interpretation of the UoLs themselves. That is, multi-functionality does not come about because of a series of homophonous UoLs, but instead because a given UoL associates with the spine in different positions and hence is enriched with different components of meaning. The i-spine consists of a grounding layer, responsible for the synchronization of the interlocutor's minds, and a response layer, responsible for the synchronization of the interaction itself. Like all layers of

structure on the spine, the response layer consists of a head position, which relates two arguments to each other by asserting whether they coincide or not. The coincidence feature is an intrinsic property of this head position which is valued by the UoLs that associate with it. The argument introduced by the response layer is the so-called *response set*, a set of elements that the interlocutors tend to. It can be indexed to the speaker or to the addressee. Initiating moves are defined by an addressee-oriented response set, as in (3)a, reaction moves are defined by a speaker-oriented response set, as in (3)b.

(3) a. Initiation: [$_{\text{RespP}}$ Resp-set$_{\text{Adr}}$ [+/-coin] [$_{\text{GroundP}}$ Utt]]
b. Reaction: [$_{\text{RespP}}$ Resp-set$_{\text{Spkr}}$ [+/-coin] [$_{\text{GroundP}}$ prev.Utt]]

In an initiating move, the content of the utterance is asserted to be or not to be in the addressee's response set. Thus, RespP allows a speaker to explicitly mark an utterance as requiring a response. In a reacting move it is the content of the preceding move by the interlocutor which is asserted to be in the speaker's response set. Thus, RespP allows a speaker to explicitly mark an utterance as being or not being a response. In what follows, I argue that these ingredients of the i-spine allow for a straightforward analysis of the adjacency pair in (1).

4 The proposal

Recall from section 1 that there are three questions the minimal interaction in (1) raises. First, what, if anything does the stare "mean" so it can trigger *what* as a response and how is this meaning derived? Second, what does *what* mean in this context, and how is its meaning derived? And finally, how do visual and verbal interaction combine to produce a coherent conversation? In this section, I introduce my answers to these three questions.

First, I propose that the stare serves as a UoL which explicitly marks an initiation move. More specifically, I propose that the stare is a complex UoL, hence its length: it is associated both with ResP and simultaneously functions as the content of the move. Thus, it places the stare itself into the addressee's response set. This is schematized in (4).

(4) [$_{\text{RespP}}$ Resp-set$_{\text{Adr}}$ [$_{\text{Resp}}$ 👁 [$_{\text{GroundP}}$ 👁]]]

Second, I argue that *what* in (1) receives its interpretation because of its intrinsic meaning (a variable restricted to inanimate individuals, which may include situations or propositions) in combination with the contribution of the response layer in the i-spine. Specifically, I argue that it associates with a complex response structure which in turn derives a move that is simultaneously a reaction and an initiation. This is schematized in (5). It can roughly be paraphrased as "Tell me, what is your initiation".

(5) [$_{\text{RespP}}$ Resp-set$_{\text{Adr}}$ [$_{\text{Resp}}$what [$_{\text{RespP}}$ Resp-set$_{\text{Spkr}}$ [$_{\text{Resp}}$~~what~~ [$_{\text{GroundP}}$]]]

It should now be clear how this analysis answers the question regarding how visual and verbal interaction combine. It is made possible because of the i-spine, which significantly contributes to the interpretation of signs and these signs need not be Saussurian signs of the familiar sort (i.e., sound-meaning bundles). It is precisely with such non-standard UoLs, which are not arbitrary

form-meaning pairings, that the contribution of the spine emerges. In what follows, I provide evidence for this analysis. I show that without the mediation via an abstract system (like the spine) it would be difficult to account for the intricate patterns that this adjacency pair displays.

5 The grammar of *what*

In this section, I explore the question as to what *what* means in (1) and how it gets to have its meaning. I present evidence that it is not used as a regular wh-word in initial position of an elided content question (section 5.1). Rather it is used to inquire about the nature of the preceding initiation move and hence is purely interactional (section 5.2).

5.1 "What" is not short for "What do you want?"

When considering the question as to what *what* means in (1), an obvious hypothesis to consider is that *what* is short for a full wh-question where the remainder of the clause is simply elided. This is a plausible hypothesis for the following reasons. First, full wh-questions with initial *what* are possible as a response to a stare, as in (6).

(6) I: 
R: a. **What**?
b. **What** are you looking at?
c. **What** do you want?
d. **What** is your problem?

Furthermore, ellipsis of this form is otherwise well-formed, as shown in (7)-(9). Here the initiation move is itself a full question, which functions as a summons. The following reaction move (which is also a question and hence simultaneously functions as a reaction and an initiating move) can be either the single-word utterance *what* (as in (1)) or a full-fledged question which repeats the question of the initiating move.

(7) I: You know what I'm looking at?
R: a. **What**?
b. **What** are you looking at?

(8) I: You know what I want?
R: a. **What**?
b. **What** do you want?

(9) I: You know what's my problem?
R: a. **What**?
b. **What** is your problem?

Thus, the question in the initiating move serves as the antecedent for the elided string in the response in (7)-(9). It is not clear what might serve as the antecedent for the hypothetical ellipsis in (1), as the preceding initiation move is just the stare. But suppose that the antecedent is somehow implicit, in a way to be made precise. The point of the data in (7)-(9) is that they show that ellipsis of this kind is a possibility.

There are however, two problems, which rule out this possibility as an analysis for the use of *what* in (1). First, consider the fact that *what* appears to be the only wh-word that can be used as a response to the stare. Crucially *why* is not possible as a reaction to a stare ((10)a) even though full *why* questions are ((10)b/c)

(10) I: 
R: a. ***Why**?
b. **Why** are you looking at me?
c. **Why** are you not saying anything

And crucially, in the presence of a full antecedent, ellipsis is possible even following *why*, as shown in (11) and (12).

(11) I: You know why I'm looking at you?
R: a. **Why**?
b. **Why** are you looking at me?

(12) I: You know why I am not saying anything?
R: a. **Why**?
b. **Why** are you not saying anything?

Given that both *what* and *why* questions are possible and given that both *what* and *why* questions license ellipsis, it is not clear why only *what* but not *why* can be used after a stare. This invites the conclusion that *what* in (1) is not an instance of an elided question.

This conclusion is supported by a second problem with the ellipsis analysis, namely that *what* is not possible in all situations even when *what* questions are. This is shown in (13), where there is clearly no initiation move and *what* in isolation is ill-formed whereas a full *what* question is possible.

(13) Upon entering a room where two people are fighting:
I: a. ***What**?
b. **What** is going on?

Note that there is nothing wrong with this particular wh-question such that it would not license ellipsis. It does when there is an appropriate antecedent in the initiating move, as in (14).

(14) I: You know what's going on?
R: a. **What**?
b. **What** is going on?

In sum, I conclude that *what* in response to a stare cannot simply be a wh-word followed by an elided full question. This is because other wh-words are not possible even though they allow for ellipsis, and second, bare *what* is not possible in all contexts even when full *what* questions are possible.

Having ruled out an ellipsis analysis for *what* in (1), I now turn to an alternative analysis according to which *what* is purely interactional.

5.2 'What' is interactional.

The hypothesis I introduce here is that the use of what in (1) is purely interactional and by this, I mean that it is not used to inquire about any propositional content as in cases of typical content questions (e.g., *What did you say*?). Instead, it is used to inquire about the nature of the preceding move. I argue that this is a result of associating *what* with the i-spine, as in (15), repeated from above. According to Author (to appear a) a move which serves simultaneously as a reaction and an initiation is complex: it consists of a speaker oriented RespP (encoding the reaction move) and an addressee oriented RespP (encoding the initiation).[1] I propose that *what* associates with both RespP and this is why it serves two functions. In the speaker-oriented RespP it serves to mark the utterance as a response, whereas in the addressee oriented Resp it serves to place the utterance into the addressee's response-set thus serving as a *go-ahead* marker.

(15) $[_{RespP}$ Resp-set$_{Adr}$ $[_{Resp}$what $[_{RespP}$ Resp-set$_{Spkr}$ $[_{Resp}$~~what~~ $[_{GroundP}$]]]

Translating this configuration into propositional language would amount to asking, *"Tell me, what is your initiation?"* That is, I propose that *what* in this context explicitly asks for the nature of the initiation. This is, however, not a matter of its lexical entry but rather of its position in the i-spine. According to this analysis, it is the i-spine which contributes explicit reference to initiation rather than aspects of propositional structure or the lexical entry itself.

This analysis predicts that *what* does in fact require the presence of an initiation move. Hence it accounts for the data in (13) above. While a propositional *what* question is possible in the absence of an initiation move, the interactional use of *what* is not.

Evidence for the claim that *what* is used as a response to a preceding initiation move comes from the fact that it can be used in reaction to all types of summons, including terms of address (*Mommy, Konrad)*, courtesy phrases (*excuse me*), attention-getting particles (*hey)*, and the stare. Crucially, all of these summonses can be responded to in the same way, including *what* as well as *yes* and full propositional *what*-questions. But even though full *why*-questions are possible as a response to these summonses, bare *why* is not. This is shown in (16).

(16)

I:	a.	Mommy/Konrad
	b.	Excuse me.
	c.	Hey.
	d.	👁
R:	a.	What
	b.	Yes/yeah?
	c.	What do you want?
	d.	Why are you calling me?
	e.	*Why?

This raises the question as to why *what* but not *why* is a possible response to a summons. Why couldn't *why,* like *what,* be associated with the i-spine to ask about the reason for the preceding initiation move? I propose that this has to do with the presuppositions associated with *why.* To see

[1]An addressee-oriented RespP cannot further be dominated: once the utterance is put into the interlocutor's response-set, the current speaker has to end their turn.

this, consider regular content questions. A *what*-question can be responded to by denying that there is something that corresponds to the variable introduced by *what.* This is shown in (17) and (18).

(17) I: **What** did you eat?
R: **Nothing.**

(18) I: **What** do you want?
R: **Nothing.**

A *why* question on the other hand presupposes that the event whose reason is being questioned has happened. This is shown in (19) and (20).

(19) I: **Why** did you eat?
R: #I didn't eat.

(20) I: **Why** are you looking at me?
R: #I'm not looking at you.

Crucially, a summon is a special kind of initiation, which may occur simply to attract the attention of the interlocutor but without conveying content (Filipi, 2009). Since *what* does not presuppose content, it is compatible with this use. In contrast, when using *why* the responder has to be sure that there is in fact an initiation and that this is shared knowledge. As is clear from the conversation in (2), it can be denied that the stare is an initiation. Hence, the existence of an initiation is not presupposed. This is why *what* is possible, but *why* is not.

Another piece of evidence that *what* is used to ask about the nature of the initiation move comes from the fact that it cannot be used as a response when it is clear what the initiation is. In other words, it is restricted to summons initiations. This is shown in (21). The initiation move here is an informative declarative statement and *what* in isolation is not possible as a response, though a full propositional *what* question is, and so is *what* preceded by *so.*

(21) I: It's raining.
R: a. *What?
b. What are you trying to say?
c. So what?

With the full *what* question and with *so what* the responder acknowledges that there is an initiation move; what is questioned here is the relevance of the content of the initiation and not the nature of the initiation itself. Note that this is not possible as a reaction to a stare, whose propositional content cannot be questioned as there is none (see below).

(22) I: 
R: a. What?
b. *What are you trying to say?
c. *So what?

We have now seen that *what* in isolation can be used in purely interactional ways, without propositional content. It serves as an inquiry about the nature of a preceding initiation move. This provides us with clear evidence that a stare can serve as an initiation move. It is not just considered an event that can be responded to; it is interpreted as an interactional event. In the next section I address the question as to how the stare gets to have this meaning in interaction.

6 The grammar of eye-gaze

In this section, I introduce in more detail the hypothesis that eye-gaze can function as a UoL which can associate with the i-spine like words. Specifically, I argue that eye-gaze can associate with the spine in RespP where it marks the utterance as an initiation turn by placing it into the interlocutor's response set. As I will show, this is a common property of eye-gaze, which has been established to play a crucial role in turn-taking (Kendon 1967). What is, however, special about the stare in the conversation in (2) is its length and the absence of an actual utterance. I propose that the length of the stare indicates actual complexity in that the eye-gaze not only serves to mark an initiation move, but it also serves as the content of the move itself. That is, in the absence of other content, the eye-gaze exceptionally plays this role by associating with the complement of Resp as well.

(23) $[_{\text{RespP}}\ \text{Resp-set}_{\text{Adr}}\ [_{\text{Resp}}$ 👁 $[_{\text{GroundP}}$ 👁 $]]]$

This analysis captures the fact that the stare functions as an initiation move, a summons. In this case, it is the eye-gaze which does both: regulate turn-taking and providing the content of the turn itself. The purpose of this section is to provide evidence for this analysis. I proceed as follows. In section 6.1, I review evidence for assuming that non-vocal signs can be associated with the spine and hence that there is no intrinsic reason not to assume that eye-gaze can function in this way. I then review the literature that shows that eye-gaze plays an important role in regulating turn-taking (section 6.2). Finally, I show in section 6.3 that eye-gaze is qualitatively different from regular vocal UoLs in that it also has a function off the spine, and this is the reason the interlocutor can deny its interactional significance and moreover that it is perceived as aggressive.

6.1 Non-vocal signs on the spine

I start with a brief overview of the literature on the grammatical status of non-vocal signs. One of the key insights behind the interactional spine hypothesis is the assumption that particles such as *oh* and *huh* which have traditionally been viewed as being outside of grammar, and which are restricted to language in interaction, are composed in the same way as conventional words with lexical content and grammatical features.

The systematicity of non-canonical UoLs has long been recognized in frameworks that explicitly explore i-language. For example, according to Clark (1996:156) "Most signals are composite signals, the artful fusion of two or more methods of signalling. […] Some might conclude that the non-linguistic methods are crude, unsystematic, ad hoc, and marginal, and deserve to be relegated to the periphery of language use. This wouldn't be right. On the contrary, the non-linguistic methods are subtle, highly systematic, and not at all ad hoc. And they are part and parcel of most signals that are usually classified as "linguistic". (See also Kendon, 1967; Bavelas and Chovil, 2000).

For Author (to appear a), the system that regulates composition is the universal spine; it equally regulates the composition of elements with propositional and interactional content. One of the pieces of evidence that i-language, too, is regulated by the spine comes from the fact that intonational contours do not distinguish between propositional language and i-language. To see this, consider the data in (24), where ↑ represents rising intonation. Sentence final-rise, which typically triggers a question interpretation (24)a, obligatorily occurs in sentence-final position. In the presence of a sentence-final interactional particle, such as *eh,* sentence-final rise has to associate with the particle (24)b. Neither can it simply associate with the propositional part of the clause, ignoring the particle (24)c, nor can it simultaneously occur on the sentence and the particle (24)d.

(24) a. You have a new dog↑
b. You have a new dog, eh↑
c. *You have a new dog↑, eh
d. *You have a new dog↑, eh↑

This suggests that for the system that regulates prosody, there is no intrinsic difference between propositional and i-language. This follows straightforwardly if they are regulated by the same system, namely the spine. In fact, the simplest way to analyse these facts is to assume that intonation itself acts as a UoL that associates with the spine (Author & Co-author 2016, Co-author & Author 2020) and hence that its distribution is regulated by the spine. If this analysis is on the right track, it implies that UoLs that do not belong to the classic set of words, morphemes, and features are composed in similar ways.

Note that the claim that intonation behaves like a UoL has been argued for before (Truckenbrodt 2012). But if it were a regular morpheme, then it would be unexpected that rising intonation is almost universally associated with a questioning interpretation. Other types of morphemes are characterized by an arbitrary relation between form and meaning. Author (to appear b) argues that the non-arbitrariness in the form-meaning relation in intonation follows from the fact that it is sound only which associates with the spine and given that the spine with its functions is universal, the meaning that intonation carries reveals the functions of the spine and is therefore not subject to the same type of variation as regular UoLs.

Other types of non-standard UoLs which have been argued to be composed syntactically are co-speech-gestures. Esipova 2018, 2019, analysing both iconic and non-iconic gestures, argues explicitly that syntax and semantics are modality blind. That is, words, gestures, and facial expressions are compose with the same system and no special semantic status has to be assumed for them (as for example in Schlenker 2018). Given that gestures clearly have a form, and a meaning, and given that these can be integrated into the overall interpretation of an utterance, assuming that the same system is responsible for doing so is the most parsimonious assumption. Esipova presents syntactic evidence to support this claim. Assuming the regulating system to be the spine, we may conclude that gestures and facial expressions can be added to the set of UoLs that can associate with the spine.

Finally, another non-verbal UoL which has recently been argued to be part of regular semantic composition is laughter, along with smiling, sighing, eye-rolling, and frowning (Ginzburg et al. 2020). Specifically, they argue that these types of signs are best integrated into a conversationally-oriented view of grammar. The assumption of the i-spine allows us to capture such elements which are restricted to language in interaction. Furthermore, given that the i-spine

dominates the spine which is used for the composition of propositions and thus truth-conditions, it follows that the meaning it contributes cannot be truth-conditional. Rather it is the natural habitat for so-called use-conditional meaning (in the sense of Gutzmann 2015), which is precisely the type of meaning that arises in interaction. And here I propose that eye-gaze, too, is part of i-language. Specifically, I propose that eye-gaze can associates with RespP and thereby functions to regulate turn-taking. In the next subsection, I review the previous literature on the role of eye-gaze in social interaction and we shall see that it has been clearly established that eye-gaze plays an important role in turn-taking in ways that are expected based on the analysis I propose.

6.2 The role of eye-gaze in turn-taking

It has long been known that eye-gaze is socially significant; Tomkins 1963 suggests that awareness of this is evident in clay-tablets from a civilization in Iraq of the Third Millennium, BC. But it was Kendon 1967, who first established that eye-gaze plays a systematic role in communicative interaction (see also Argyle and Cook 1976 and Goodwin 1980, 1981). Based on analyses of videos of conversations in controlled experimental settings, and through detailed analysis of the relation between eye-gaze and utterances in conversation, Kendon found that eye-gaze correlates with turns. Specifically, at the end of a turn, speakers gaze at their interlocutors, while at the beginning of a turn, speakers tend to look away. Based on these findings, Kendon suggests that the turn-final eye-gaze has both a monitoring and a regulating function. By looking at their interlocutor, the speaker signals to their interlocutor that they are ready to end their turn and hence that they are ready to listen to their interlocutor's turn. At the same time, the gaze functions to monitor whether this signal has been received and whether indeed the interlocutor is ready for their turn. For the turn-initial gaze aversion, Kendon suggests that it signals that the interlocutor has accepted the offer of change of role: they now hold the turn. Gaze aversion does not have an obvious monitoring function in the same way as mutual gaze does. That is, by looking away, the new turn-holder does not monitor their interlocutor's behavior, because monitoring something requires looking at it.

Kendon's early findings and his interpretation of them have more or less stood the test of time. It is by now uncontroversial that gaze is not only used to have visual clues regarding the attention of one's interlocutor, but that it is also communicative (Wu et al. 2014,) and moreover that there is a systematic corelation between gaze and speaking (Ho et al. 2015). In a recent review of methods and findings in the study of the role of eye-gaze in turn-taking, Degutyte and Astell 2021 find that for turn-final gazes, which facilitate turn-yielding, the results have been replicated again and again in experimental and natural settings, with sophisticated methods including eye-tracking. However, when it comes to eye-gaze at the start of a turn, the results are more variable. While the observation that speakers tend to look away at the beginning of a turn has been replicated (Duncan & Fiske 1977, Cummins 2012), its interpretation varies. For example, Ho et al. (2015: 15) find that speakers tend to begin their turn with averted gaze and gazing at one's interlocutor occurs only around 700ms after the beginning of the turn. They suggest that aversion may "signal a desire to maintain the turn, letting the partner know that they have the floor." According to another study, gaze aversion at the beginning of a turn results from the high cognitive load of the task at hand (Doherty-Sneddon & Phelps 2005). Looking at someone's face (direct eye-gaze) is highly informative and hence requires cognitive resources (Glenberg et al., 1998). Thus, gaze-aversion frees cognitive resources to be used for the turn itself. If so, then gaze aversion would not intentionally be used to signal anything about turn-initiation, even though it might correlate with it, at least sometimes. Note that gaze aversion outside of an ongoing conversation may

have the opposite effect. That is, Goffmann 1963 argues that in some contexts, unacquainted individuals owe one another a brief acknowledgement of presence via eye-contact (e.g., when entering into a waiting room), but as a rule this is followed by a withdrawal of attention to indicate that the other is of no special concern. It is precisely when visual attention proceeds for longer than required that an unacquainted person can signal a desire for conversation. Cary 1978 establishes experimentally that the initiation of a conversation among unacquainted people can be predicted based on the amount of eye-contact that takes place even before the conversation. Consequently, we have to conclude that gaze aversion is not necessary for the start of a turn. In some contexts, the converse is true: mutual gaze is necessary to start a conversation. Thus, prolonged eye-gaze can function to indicate the desire to initiate a turn and thus may function as a summon. Though typically summons are accompanied by verbal clues (see section 2). What is crucial for the purpose of our discussion is the fact that summonses are initiating turns, yet they are not characterized by gaze aversion. Rather the eye-gaze has exactly the kind of functions ascribed to turn yielding: it regulates turn-taking by signalling that the current interlocutor wishes to hand over the turn and it is also used to monitor whether the addressee is ready to take that turn.

Based on the gaze properties of summons, we might conclude that gaze aversion characterizes *reaction* and therefore it only appears to signal the beginning of a turn. It is because it signals reaction to a previous turn that it occurs turn-initially; it does not signal turn-initiation as we can conclude from the fact that in first move initiations it does not occur.

In sum, it is uncontroversial that eye-gaze plays a crucial role in regulating turn-taking, and that it does so in systematic ways. This suggests that there is an underlying system that regulates this form-function pairing. What I suggest here is that this underlying system is the i-spine: eye-gaze can function as a UoL and associate with the spine. The advantage of this assumption is that it allows us to shed new light on the multi-functionality of eye-gaze. That is, it is clear that the regulation of turn-taking is not an intrinsic part of the "meaning" of eye-gaze. We know this for at least two reasons. On the one hand turn-taking occurs in the absence of eye gaze (e.g., when talking on the phone) and hence the regulation of turn-taking cannot depend on eye-gaze; it exists independent of the form it takes in the presence of eye-gaze. This follows because the functions associated with the i-spine do not depend on their expression via eye-gaze. On the other hand, as we have seen, eye-gaze is not always used to regulate turn-taking but has other functions as well (such as monitoring someone or something to gather visual evidence). Thus, regulating turn-taking cannot be intrinsically associated with eye-gaze. It is only when eye-gaze associates with the i-spine that it acquires the function of regulating turn-taking.

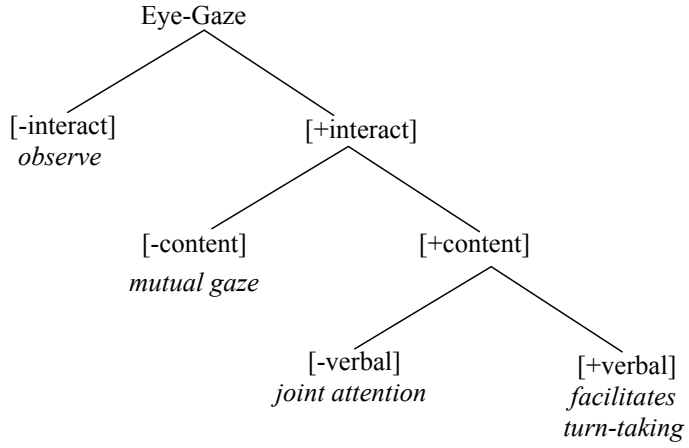
6.3 The special status of the stare summons

We have now seen that eye-gaze, like other UoLs, displays patterns of multi-functionality. I argue that this multi-functionality derives at least in part from its association with the spine. But crucially, there is a qualitative difference between the multi-functionality of eye-gaze and that of other UoLs. This has to do with the fact that eye-gaze exists outside of conversations, and even outside of interaction. That is, eye-gaze is what happens when we look at things and we look at things for most of our waking hours. This is not the case for standard UoLs, like words, or even non-standard UoLs, like intonation or gestures. Words, unlike eye-gaze, are intrinsically tied to language: in the absence of language there are no words. Trivially, words cannot be used outside of verbal

interaction. For intonation and gestures, however this is not true: roots of intonation are detectable outside of the use of language. For example, it has been shown that the melody of infant's cries is shaped by their mother's language. (Mampe et al. 2009). Assuming that the melody of a cry is a form of intonation might indeed indicate that the use of intonation is independent of the use of language. As for gestures, they, too, appear in the absence of language: for example, apes are known to use gestures for communicative interactions though they do not have language (Pollick & de Waal 2007). Similarly, children use gestures before they use language. And adults can use gestures without language and even without communicating with others, as in self-talk, for example. Though one could argue that even self-talk is a form of communication.

Eye-gaze however differs in that it is ubiquitous, unlike intonation and gesture, which are special events. Rather eye-gaze is necessary for visual perception, which in turn is unequivocally independent of interaction, let alone verbal interaction. In sum, what makes eye-gaze special is that it has purely perceptual functions, but it simultaneously plays an important role in interaction, both verbal and non-verbal. When we see things, we can't be said to interact with those things, but when we see another human being and especially when we look them in the eyes (mutual gaze) we interact. This kind of interaction, however, is not about communicating something, it is, in a way, without content, i.e., the interaction is not *about* something, it simply consists of looking each other in the eyes. But eye-gaze can also be used to communicate about something; this is what happens in so-called *joint attention*. Joint attention is the ability to intentionally co-orient towards a common focus (Leavens & Racine 2009) and it can be achieved via explicit (gestural) pointing but also via eye-gaze alone. The use of eye-gaze in regulating turn-taking is thus special in that it correlates with verbal interaction, which eye-gaze itself does not need. The multi-functionality of eye-gaze is summarized in (25).

(25) The multi-functionality of eye-gaze



With this in mind, let us turn back to the conversation introduced in section 2 and repeated below as (26).

(26)
01 E ((looks up at T))
02 (3.2)
03 T→ What.
04 (2.2)

05 E Nothing. I didn't
06 say anything.
07 T °Don't stare at
08 me then°
09 J St[op being so=
10 E [What?
11 J =aggressive Emily
12 E Actually Tom didn't
13 actually need to…

Clift: F1:6:24:19',

The fact that eye-gaze is used outside of verbal interaction, and even outside of social interaction more generally is the reason that E may deny her initiation while T's response clearly indicates that he interprets her stare as an initiation. That is, as a response to T's *what*, E denies any initiation. It is of course true, that E didn't *say* anything, but this does not imply that her behaviour cannot be construed as an initiation. Thus, while T interprets E's stare as an initiation move (and, as we have seen he is justified to do so), E is equally justified to deny the existence of such a move, because she may have just looked without initiating a communicative act. Note that verbal summonses are not ambiguous in this way, as shown in (27). A verbal summons, can be answered with *what* but in this case the initiator cannot not deny the initiation. It simply is not true that they didn't say anything.

(27) I: Hey Kelly!
R: What?
I: * Nothing, I didn't say anything.

Once E denies initiation, this gives rise to T's request not to stare. This, too, is fully justified as a next move in light of the fact that E just denied initiation. E cannot deny that she stared at T but staring at other people has profound effects and as a rule is to be avoided. This is true for the unacquainted in public spaces (cf. Goffmann's 1963 social norm of *civil inattention*). Strangers are meant to show appropriate amount of indifference to one another, otherwise stares can be perceived as a threat. This is evidenced, for example, by the so-called *watchful eye effect:* feeling watched likely elicits negative emotions (e.g., anxiety, distress, nervousness, Panagopoulos & van der Linden 2017). And humans (like many animals) are especially equipped to perceive other's eye-gaze, presumably as a mechanism to detect danger (Haxby et al. 2002). Moreover, it has long been known that stares in humans and animals alike can be used to display and establish dominance (Lawless 1976) and thus can be perceived as a threat signal (Nichols & Champness 1971). As a consequence, many animals respond to direct gaze with displays of fear, aggression or submission (Schwab & Huber 2006). Thus, the aggression E is accused of by her mother is real and deeply rooted in social cognition, much deeper than language is.

7 Conclusion

7.1 Summary

The goal of this paper was to analyse in detail a single, minimal, conversation, consisting of a stare and *what* as a response. At first sight, this appears to be a straightforward summons-answer pair. The original contribution of this paper is to address the question as to how the forms used are associated with the function they have. Specifically, how can a stare function as a summons? And what exactly does *what* mean and how? And how can we understand the multi-modal character of this conversation, i.e., how do visual and verbal interaction combine to produce a coherent conversation? Within pragmatic analyses and conversation analysis, the question regarding form-meaning pairings is typically not addressed and within the generative enterprise, where the relation between form and meaning is at the core, conversations are not the unit of analysis, and non-verbal UoLs are typically not considered. Traditionally, both phenomena have been considered to lie outside of language competence, namely in the realm of performance. However, as established in conversation analysis, and other frameworks that deal with language in interaction, language competence is not restricted to sentences in isolation but extends to pragmatic knowledge.

To answer these questions, I have used the interactional spine hypothesis, introduced in Author (to appear a). This framework seeks to model the relation between form and meaning in the language in interaction, including non-canonical UoLs. The key advantage of a grammatical analysis of this type is the fact that the i-spine allows us to understand how a particular UoL (such as *what*) can be enriched to have a complex meaning. It allows us to model the multi-functionality of UoLs: depending on their position along the spine, they are interpreted in different ways because the spine contributes meaning to UoLs. In the absence of the meaning provided by the spine, one would have to postulate several homophonous lexical entries. However, the systematicity of patterns of multi-functionality casts doubt on any approach that postulates accidental homophony. Furthermore, the i-spine allows us to model the use of non-canonical UoLs, such as gestures and eye-gaze. They can function as UoLs in that they, too, can associate with the spine and hence are expected to be enriched with the same functions as regular UoLs. But they are also predictably different. Since they are not arbitrary bundles of sound and meaning (i.e., classic Saussurian signs), we expect that their meaning is composed somewhat differently, and indeed it is. Thus, the advantage of the i-spine is that it straightforwardly allows for the integration of multi-modal channels. The present paper is a first attempt to integrate the role of eye-gaze into grammatical analysis. It invites several avenues of future research, some of which I briefly outline below.

7.2 Future research.

Since multimodality in conversations is a novel empirical domain for formal grammatical analysis, it requires a novel methodology. Traditionally, grammatical analysis of the generative kind relies on native-speaker judgements of well-formedness for data-collection. And despite criticism, this methodology has stood the test of time (Sprouse and Almeida 2012, 2018). In fact, it sometimes is the only way to elicit minimal pairs as well as negative data. It contrasts with the traditional methodology of conversation analysis, which is corpus analysis. What I am suggesting is not to abandon corpus analysis altogether. Rather I submit that a combined approach is necessary (cf. Clark and Bangerter 2004). This is what I tried to achieve in this paper: the starting point was a data-point found in a corpus along with a qualitative analysis. But many of the data that led to the hypothesis I put forward for *what* came from subsequent native speaker judgments. But what remains to be developed is a protocol for elicitation of non-canonical UoLs, such as eye-gaze.

Another avenue for future research has to do with patterns of markedness. That is, according to the universal spine hypothesis (Author 2014), upon which the interactional spine

hypothesis is based, the linguistic behaviour of categorization of UoLs is characterized by two properties: patterns of multi-functionality and patterns of markedness. Here we have seen that patterns of multi-functionality are indeed found in eye-gaze. But what remains to be seen is whether patterns of contrast, too, can be found. That is, for any category it is typically the case that there are (at least) two values associated with it. This follows form the assumption that the spine comes with an intrinsic coincidence feature, which is valued by the UoL that associates with it. In essence, the head in combination with the UoL will assert whether or not the two arguments that are being related coincide or not. We here have seen that eye-gaze leads to a request for a response (turn-yielding). We might expect then that gaze-aversion as found in reactions may be the contrasting value, at least in certain contexts: it signals the opposite of a request for response. It indicates that the speaker wishes to keep their turn.

Finally, the present approach makes it possible to explore the form and function of eye-gaze from a cross-linguistic perspective. Given that eye-gaze is a UoL which is not characterized by an arbitrary form-meaning relation we expect there to be universal patterns of function: this is because the gaze itself is associated with particular functions and the i-spine is by hypothesis universal.

Finally, given the assumption that the communicative function of eye-gaze is dependent on the spine, we predict that eye-gaze will have different properties in individuals whose language faculty (and hence the spine) differs from typical developing adults. Specifically, we predict that eye-gaze will not correlate with turn-taking, at least not in the same way. This prediction can be explored based the role of eye-gaze in language development, clinical populations (i.e., individuals that lack the capacity for language). Finally, we also expect that animals that display communicative behaviour which includes some form of turn-taking will also lack the particular properties of eye-gaze characteristic of humans. And more specifically, we predict that a stare cannot be interpreted as a summons. I take the fact that the proposed analysis makes predictions of this type as a virtue; it invites a novel research approach towards the function of eye-gaze in human communication and beyond.

8 References

Wiltschko, M. to appear. *The universal structure of categories. Towards a formal typology.* Cambridge University Press.

Wiltschko, M. to appear a. *The grammar of interactional language.* Cambridge University Press.

Wiltschko, M. to appear b.

Author and Co-author. 2016.

Co-author and Author. 2020.

Argyle, M. & Cook, M. 1976. *Gaze and mutual gaze.* Cambridge University Press.

Bavelas, J. Beavin, & N. Chovil 2000. Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology* 19(2): 163-194.

Cary, M. S. 1978. The role of gaze in the initiation of conversation. *Social Psychology* 41(3): 269-271.

Clark, H. 1996. *Using Language.* Cambridge University Press.

Clark, H., & A. Bangerter. 2004. Changing ideas about reference. In: I. Noveck & D. Sperber. (eds.) *Experimental Pragmatics*. Palgrave Studies in Pragmatics, Language and Cognition. Palgrave Macmillan, London. 25–49. https://doi.org/10.1057/9780230524125_2

Cummins, F. 2012. Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes* 27:1525–49. doi: 10.1080/01690965.2011.615220

Degutyte, Z. & A. Astell. 2021. The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, 12. https://doi.org/10.3389/fpsyg.2021.616471

Doherty-Sneddon G, & F. Phelps. 2005. Gaze aversion: a response to cognitive or social difficulty? *Memory and Cognition* 33:727–33.

Duncan S & D. Fiske. 1977. *Face-to-face interaction: Research, methods, and theory*. New Jersey: L. Erlbaum Associates.

Esipova, M. 2018. Focus on what's not at issue: Gestures, presuppositions, supplements under contrastive focus. *Proceedings of Sinn und Bedeutung* 22: 385–402.

Esipova, M. 2019. *Composition and projection in speech and gesture*. Doctoral Dissertation, NYU.

Filipi, A. 2009. *Toddler and Parent Interaction: The Organization of Gaze, Pointing and Vocalization.* Amsterdam: John Benjamins.

Ginzburg, J., C. Mazzoconi & T. Ye. 2020. Laughter as language. *Glossa*: *a journal of general linguistics* 5(1): 104. 1–51. DOI: https://doi.org/10.5334/gjgl.1152

Glenberg, A., J. Schroeder & D.Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition* **26,** 651–658. https://doi.org/10.3758/BF03211385

Goffman, E. 1963. *Behaviour in public places. Notes on the social organizations of gatherings*. New York: The Free Press.

Goodwin, C. 1980. Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological inquiry*, 50(3-4), 272-302.

Goodwin, C. 1981. *Conversational organization*. *Interaction between speakers and hearers*. New York: Academic.

Gutzmann, D. 2015. *Use-Conditional Meaning: Studies in Multidimensional Semantics.* Oxford University Press

Haxby, J., E. Hoffman, & M.I. Gobbini. 2002. Human neural systems for face recognition and social communication. *Biological Psychiatry,* 51(1): 59-67.

Ho S., T. Foulsham & A. Kingstone. 2015. Speaking and Listening with the Eyes: Gaze Signalling during Dyadic Interactions. PLoS ONE 10(8). doi:10.1371/journal.pone.0136905

Jefferson, G. 1972. Side sequences. In: D.Sudnow (ed.) *Studies in social interaction.* New York: Free Press. 294-338.

Ochs Keenan, E. & B. Schieffelin. 1976. Topic as a discourse notion: A study of topic in the conversations of children and adults. In: Li C. (ed.) *Subject and Topic*. New York Academic Press. 335-384.

Kendon, A. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, *26*, 22-63.

Lawless, W. 1976. The Role of Dominance in Human Responses to Staring. *LSU Historical Dissertations and Theses* 2971. https://digitalcommons.lsu.edu/gradschool_disstheses/2971

Leavens, D., & Racine, T. P. 2009. Joint attention in apes and humans: Are humans unique? *Journal of Consciousness Studies*, 16(6-7), 240-267.

Mampe, B., A. Friederici, A. Christophe, & K. Werm. 2009. Newborns' Cry Melody Is Shaped by Their Native Language *Current Biology* 19, 1994–1997. DOI 10.1016/j.cub.2009.09.

McTear, M. 1985. *Children's Conversation* Basil Blackwell, Oxford.
Mondada, L. 2019. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. *Journal of Pragmatics* 145, 47-62.
Nichols, K. & Champness, B. 1971. Eye gaze and the GSR. *Journal of Experimental Social Psychology* 7(6), 623-626.
Ochs, E., B. Schieffelin, B. & Platt, M. L. 1979. Propositions across Utterances and Speakers. In E. Ochs, & B. Schieffelin (eds.) *Developmental Pragmatics*. London: Academic Press. 251-268.
Panagopoulos, C., & S. van der Linden. 2017. The feeling of being watched: Do eye cues elicit negative affect? *North American Journal of Psychology* 19(1): 113-121.
Pollick, A. & F. de Waal. 2007. Ape gestures and language evolution PNAS 104 (19) 8184-8189; https://doi.org/10.1073/pnas.0702624104
Sacks, H. 1995. *Lectures on Conversation: Volumes I & II,* Oxford: Blackwell.
Schegloff, E. 1968. Sequencing in Conversational Openings. *American Anthropologist* 70(6), 1075-1095.
Schegloff, E. A. 2007. *Sequence organization in interaction: A primer in conversation analysis.* Cambridge University press.
Schlenker, P. 2018. Gesture projection and cosuppositions. *Linguistics and Philosophy* 41:295–365.
Schwab, C., & Huber, L. 2006. Obey or not obey? Dogs (Canis familiaris) behave differently in response to attentional states of their owners. *Journal of Comparative Psychology* 120(3), 169.
Sprouse, J.& D. Almeida. 2018. Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable.*Behavioral and Brain Sciences.* 40: e311.
Sprouse, J. & D. Almeida 2012. Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics.* 48: 609-652.
Tomkins, S. 1963. *Affect imagery consciousness: Volume II: The negative affects*. Springer Publishing Company.
Truckenbrodt, H. 2012. Semantics of Intonation. In: C. Maienborn, K.Heusinger and P. Portner (eds.) semantics: *International Handbook of Natural Language Meaning*. NY: Mouton de Gruyter. 2039-2069
Wu, D., W. Bischof, & A. Kingstone. (2014). Natural gaze signaling in a social context. *Evolution and Human Behavior* 35(3), 211-218.