Abstract

We argue that contemporary formal grammar is unable to provide analyses for language as it occurs in actual spoken interaction. Its analyses are developed for a cleaned up version of language which omits the disfluencies, fragments, gestures, and many other phenomena that are ubiquitous in spoken language. These we show, using evidence from linguistics, conversation analysis, multimodal communication, psychology, language acquisition, and neuroscience, are rule governed in much the same way as phenomena captured by standard grammars. We argue that over the past few years the theoretical tools required to provide a precise characterizations of such phenomena have begun to
emerge in theoretical and computational linguistics. These form a coherent framework, that we will call *Interaction Grammar*; we sketch in the paper the main aspects of this framework. Finally, we suggest that a framework of this kind would provide a better foundation not just for linguistic analysis of face-to-face interaction, but also for sister disciplines concerned with such phenomena, such as research on spoken dialogue systems and psychological work on dialogue phenomena, including in particular language acquisition.

**Keywords**

Interaction and the competence/performance distinction
Interaction Grammar
Semantics of Dialogues
Non-sentential utterances
Self-repair and other-repair
References to the Interaction Situation
Quotation
Gestures and Multimodal Grammar
1 Introduction

What should grammars characterize?

For millennia grammars were developed with written language in mind. Providing analyses for examples from written texts was the standard task for grammarians. But following Saussure and the American structuralists, inter alii, spoken language became a reputable object of study as well. This trend should have strengthened with the rise of generative grammar, whose avowed aim was characterizing linguistic competence, a notion relating the ability to use spoken language that applies equally to individuals speaking English, French or Chinese in post-industrial societies and tribesmen speaking Pirahã in the Amazon or Arapesh in Papua New Guinea.

And yet, in practice contemporary grammar is not, for the most part, interested in or able to provide analyses for language as it occurs in actual spoken interaction. Its analyses are developed for a cleaned up version of language (e.g., (1b) from (1a)), which omits the disfluencies, parentheticals, fragments, and ad hoc coinages which are ubiquitous in spoken language, as exemplified in (1)-(3):

(1) a. I’m just really anxious—not anxious, anxious is the wrong word, I’m excited about tomorrow. (Roy Hodgson, England Football manager, The Guardian, 10 October, 2013)
   
   b. I’m excited about tomorrow.

(2) 1. A: and they took a bit of my bone away, also in the process, cos it was so like crck crck
   2. B: what did they put there instead?
   3. A: didn’t put anything!
   4. A: it was a huge, it was a big hole it was
   5. B: what’s there now?
   6. A: huh?
   7. B: what is what’s in your mouth now?
8. A: there’s nothing, I have like this this

9. A: this piece of gum that's, that you know erm it’s just sort of gummed back together

(From the corpus described in (Healey et al, 2015))

(3) 1. Fri: They still haven’t figured out,(.) how they’re gonna get to the country: ¡ who’s gonna take care of huh m:othah while [they’re- y’know ’p in the country. on the weekend.

2. Dav: [Mm (0.2 secs)

3. Fri: So: (. ) you know, (0.8 secs)

4. Fri: an besides th[.t,

5. Rub: [You c’n go any[w

6. Dav: [Don- Don git- don [get]

7. Fri: [they] wont be:

8. Dav: Yknow there- theres no- no long explanation is necessary

9. Fri: Oh noon no: I’m not- I jus: : uh-wanted: you to know that you can go up anyway.=

10. Rub: =Yeah:. (0.1 secs)

11. Fri: You know. (0.2 secs)

12. Fri: Because-ah (3.3 secs)

13. Rub: They don mind honey they’re jus not gonna talk to us ever again.=

14. Dav: = (laughter) / ri:(h)ight)

(From (Schegloff, 2001))

This written language bias (Linell, 2005) characterizes work in most areas of formal grammar, be it the Minimalist Program (Chomsky, 1995), Head-driven Phrase Structure Grammar (HPSG) (Pollard & Sag, 1994; Ginzburg & Sag, 2000; Sag, Wasow, & Bender, 2003), Lexical Functional Grammar (Bresnan, 1982, 2001), Categorial Grammar (Moortgat, 1997; Steedman, 2001), Construction Grammar (Goldberg, 1995; Kay & Fillmore, 1999), though, as we shall see, there has certainly been work in some of these frameworks that very directly engages spoken language.
For some frameworks the bias is explicitly justified, given continued adherence to Chomsky’s competence/performance distinction (Chomsky, 1965) and to a view of grammar as ‘the capacity for unbounded composition of various linguistic objects into complex structures . . . This approach distinguishes the biological capacity for language from its many possible functions, such as communication or internal thought’ (Hauser et al., 2014). Accordingly, some grammarians attempted to delineate core phenomena the grammar needs to account for, in contrast to a periphery (Chomsky, 1981). However, this strategy seems to have little independent justification (Jackendoff, 2005). Another strategy is to cleanly separate processing within the sentence and discourse–oriented processing, see e.g., (Frazier & Clifton, 2005). But such a strategy, whatever its merits, is not helpful for dealing with various pervasive within–sentence conversational phenomena such as disfluencies and interjections, exemplified above and discussed below. In other cases, less commitment has been explicitly made as to what empirical phenomena grammars needs to account for.

Our main claim in this paper is that grammatical competence requires making reference to conversational (and more generally) social interaction— it cannot be described in interaction–free terms. Just like physics takes responsibility for explaining all physical phenomena (and does not e.g., restrict itself to frictionless abstractions) and biologists do not put to one side duck billed platypuses or non–kin oriented altruism, grammars need to aim to analyze all aspects of language use. We will argue that phenomena such as fragments, references of various form to the Interaction Situation, multimodal utterances, and even disfluencies are subject to grammatical constraints of various types. In other words, we subscribe to Chomsky’s claim that ‘The behavior of the speaker, listener, and learner of language constitutes, of course, the actual data for any study of language.’ (Chomsky, 1959). However, these phenomena cannot be explained within ‘standard’–interaction-free, as we will call them–conceptions of grammar, which are therefore intrinsically incomplete. One needs grammars that can encode a view of compositionality
wherein meaning emerges by combining information from the Interaction Situation, speech events, and gestures.

Over the past 40 years there has been important work by researchers in conversation analysis, cognitive and social psychology (e.g., (Hymes, 1972; Allwood, 1976; Schegloff, Jefferson, & Sacks, 1977; Levelt, 1993; Clark, 1996; Pickering & Garrod, 2004; Linell, 2009) ). From this research, has come a clear awareness that competence needs to be stated within a conversationally oriented view of language (see also (Ono & Thompson, 1995)). While this research has yielded many important insights, some of which are mentioned below, it has for the most part not been formulated within formal frameworks of grammar or cognition. Nor has it developed a precise theory of the structure and dynamics of context in conversation. This has allowed the impression to be conveyed that the various phenomena uncovered in this research cannot or should not be described within theories of grammar similar to those used to describing the more traditional ‘cleaned up’ grammatical phenomena.

The second main contention of this paper is that over the past few years a theoretical framework has begun to emerge in theoretical and computational linguistics in which such precise descriptions can be provided. We use the term framework because what we are referring to is not yet a fully-specified theory, but rather a collection of hypotheses about the grammar of language use in interaction and of theoretical tools developed in a number of distinct theories of particular aspects of language use e.g., (Ginzburg, 2012; Purver, Cann, & Kempson, 2006; Poesio & Rieser, 2010). This collection is nevertheless coherent enough, in our view, to deserve the name Interaction Grammar that we will use.

The structure of the paper is as follows. In Section 2 we briefly discuss some phenomena whose meaning turns out to be intrinsically interactive and which modern grammatical frameworks have treated, though typically without interfacing with a detailed treatment
of context. In Section 3 we present linguistic evidence supporting the contention that interaction–free grammar will not work for spoken language, based on an analysis of a wide variety of ubiquitous constructions. In Section 4, we present evidence supporting the interaction-oriented view of grammar from other disciplines that study language, specifically language acquisition and cognitive neuroscience. Section 5 presents the aspects of interaction grammar that are beginning to emerge from various theoretical proposals; in particular, it sketches an account of all the phenomena discussed in Section 3. Finally, in Section 6 we briefly discuss the implications of our proposal for linguistics and other behavioral sciences.

2 Interactional aspects of communication already accepted as part of grammatical competence

It is worth stressing that modern linguistic theory already accepts that grammatical competence governs ways of communicating information that are only encountered in interaction, from intonation to gestures, and in particular purely gestural forms of communication such as sign language. In addition, it is already accepted that grammar governs aspects of language whose characterization by necessity involves reference to the context in which the interaction takes place, for which we use the term Interaction Situation\(^1\) (see Sections 3 and 5). The most obvious such aspect is **deictic reference**.

2.1 Intonation

It has long been accepted that (some components of) intonation needs to be incorporated in sentence level grammar and interact with the meaning introduced by words and phrases (see e.g., (Sgall, Hajičová, & Benešová, 1973; Jackendoff, 1972; Krifka, 1992; Rooth, 1993; Erteschik-Shir, 2007) among many). Crucially, some of the meanings conveyed by intonational meaning seem to be irreducibly interaction oriented—the fall-rise intonation contour (the sequence of tones L(ow)H(igh) in autosegmental theory)
associated with theme/ground in English is explicated as, roughly speaking, presupposing a certain issue being *under discussion*, whereas the nuclear pitch accent associated with focus/rheme (the high tone H) as introducing information *new for the addressee* (see e.g., (Roberts, 1996; Steedman, 2014) for detailed accounts); similarly, the French non-falling contours (a sequence ending with an H*) is used when the message conveyed is assumed to involve *controversy between speaker and addressee* (Beyssade & Marandin, 2007).

There is significant evidence of languages which express similar meanings via word order (e.g., Catalan (Vallduví, 1992), Greek (Alexopoulou & Kolliakou, 2002)) meaning that word order is also implicated interactionally. There are various attempts to integrate such notions into most modern grammatical frameworks (HPSG (Engdahl & Vallduví, 1996), LFG (Dalrymple, Mycock, Butt, & King, 2011), Minimalism (Zubizarreta, 1998)). For the most part, these do not interface with representations of context, but see (Steedman, 2014) for such an account with Combinatory Categorial Grammar and (Vallduví, 2015) for a detailed account of all notions of information structure cast in terms of dialogical context.

### 2.2 Gestures

In face-to-face conversation, verbal information is integrated with considerable information from gestures (Kendon, 1980; McNeill, 1992; Bavelas & Chovil, 2000; Kendon, 2004). And indeed, gestures were seen in traditional rhetoric as a key component of human utterance and public performance. However, gestures lost their status sometime during the nineteenth century— in part because of a shift towards a more controlled style of public delivery in which gestures played less of a role, in part because the printed word came to be see as the truest form of language expression (Kendon, 2004). This decline in status of gestures was paralleled by a reduced interest in this form of expression in linguistics. Linguists came to question the extent to which the contribution of gestures ought to be considered part of grammar, arguing instead that the role of gestures is purely
depictive or pantomimic (Kendon, 2004).

In the last thirty years, however, gestures have come to be recognized again as a key component of human utterance. Recent technological advances in recording and analyzing videos have enabled extensive and detailed empirical investigations. Studies of the relation of gesture and speech using such audio-visual methodology have shown the two activities to be so intimately correlated, that they appear to be governed by a single process (Kendon, 2004). Research e.g. in the ToGoG project provided evidence that a number of gestures have undergone a process of grammaticalization (Bressem & Ladewig, 2011; Schoonjans, 2013). There is also psychological evidence that such information is immediately integrated with information coming from speech (e.g., (Ozyurek et al., 2007)). It has also become clear that both gesture and speech make essential contributions to referential meaning, so that one form of expression cannot be considered as primary (Kendon, 2004).

One example is head-shaking and other gestures used to express negation (Kendon, 2002). A formal treatment of gestural negation and its grammatical role—in particular, its scope—has been provided by, e.g., (Harrison, 2010).

Such evidence led to the development of so-called multimodal grammars which provide an integrated account of both the spoken and the gestural aspect of human utterance (Johnston et al., 1997; Lascarides & Stone, 2009; Poesio & Rieser, 2009; Alahverdzhieva & Lascarides, 2010; Fricke, 2013).

2.3 Deixis

One type of reference to the Interaction Situation that is generally seen as governed by grammar is the information coming from pointing. In the account of demonstratives of (Kaplan, 1978), for instance, every demonstrative d is accompanied by a demonstration δ—e.g., a pointing gesture—and the grammar provides a semantics for d[δ] jointly:
specifically, $d[\delta]$ is a directly referential term that designates the **demonstratum** of $\delta$ in context $c$. This account has been widely adopted in modern formal semantics. But the idea that the role of the Interaction Situation in the semantics of demonstratives could be entirely abstracted away through the notion of demonstration is open to significant challenge, as we discuss in section 3.5.

**2.4 Sign language**

Virtually all theoretical linguists view sign language as being governed by the same kind of grammar that governs verbal forms of communication (Newport & Supalla, 1999). Accounts of, e.g., the grammar of pronominal anaphora, or the tense system of several sign languages have been proposed, that utilise the same ingredients of standard generative grammar (see, e.g., (Zucchi, 2012)).

Like the accounts of the grammatical role of gestures discussed above, such accounts abstract away from references to the Interaction Situation. But much the same issues arise with such abstraction as with the abstraction proposed for the role of pointing gestures in deixis. Indeed, the exact same issues arise for the proposed accounts of **anaphoric reference in sign language**.

Anaphoric pronouns are usually expressed in sign language by pointing to the spatial locations where the antecedents have been signed. For example, while in English sentence (4) below (Lillo-Martin & Klima, 1991) the relation the pronouns *he* and *him* bear to their antecedents is not overtly marked and needs to be inferred from extra-linguistic clues, in American Sign Language (ASL) the corresponding sentence is disambiguated by the **loci** of the pronouns: the locations in space to which the index finger points. If the index points to the location where the sign JOHN was signed, then JOHN is the antecedent of the pronoun, while if the index points to the location where the sign BILL was signed, BILL is the antecedent of the pronoun.
(4) John called Bill a Republican and then he insulted him.

Clearly, the same questions raised with respect to pointing apply to the case of loci identification.

2.5 Scaling up

Below we argue that there is no principled dividing line between phenomena such as intonation and deixis, widely accepted as falling within the purview of (competence) grammar and the phenomena we discuss henceforth. Given the need to accommodate the former within grammar, this entails a similar conclusion for the latter. This, in turn, requires a view of grammar embedded in interaction, a move which will also lead to a more principled account of ‘information structure’ phenomena.
3 Linguistic evidence about language and interaction

In this section we demonstrate that many pervasive aspects of spoken and written language use are subject to grammatical constraints that cannot be described in interaction-free terms. The phenomena we discuss include:

1. **Grammar across conversational turns**: parallelism constraints on multiple linguistic levels whose scope ranges across participant turns.

2. **Interaction Situation reference**: the existence of systematic, conventionalized dependencies that make explicit, unavoidable reference to the Interaction Situation.

3. **Gesture interactivity**: many gesture types have meanings that are intrinsically interactive.

4. **Online repair**: repair phenomena that take place while the utterance is in progress and lead to non-monotonic effects in structure and content construction.

5. **Genre dependence**: the impossibility of maintaining a global grammar.

3.1 Grammar across turns

In a wide variety of languages there exist words and phrases whose conventional meaning requires making non-eliminable reference to the existence of a conversation, indeed to the fine structure of a conversation. **Greeting** words like English *hi, hello,* *good morning* must occur conversation initially or as responses to an immediately prior greeting by another conversationalist. Some languages have more fine grained systems, e.g., Palestinian and Lebanese Arabic where *sabah. elseyr, marhaba,* and *bonjour* occur conversation initially, whereas *sabah. elnur, marhabteyn,* and *bonjouryn* can only be used as responses to these greetings, respectively.
These facts, which require direct reference to conversational structure, need to be registered in some way in the lexical entries of such words. Thus, very similar argumentation to that used by syntacticians to motivate various notions of (intrasentential) syntactic dependence e.g. cliticisation and complementation could be used to motivate the need for a mechanism that can capture the fact that words like sabah. elnur and marh.abteyn can only be used as responses to greetings of a certain form.

By the same token, a wide variety of languages have words and phrases whose conventional meaning involves parting. Parting is more complex than greeting — it involves making the judgment that a non-negligible amount of interaction has taken place (Ginzburg, 2012). As with greetings, there exist languages where the parting expression has presuppositions about the form of a preceding parting phrase: in standard Arabic Allah ya‘afik requires as preceding utterance the parting phrase ya‘atik el’afiye. This indicates that such form–oriented cross-turn presuppositions apply to multi word expressions as well:³

(6) (#) A: ya‘atik el’afiye. (‘[God] give you health’)  
      B: Allah ya‘afik (‘God healthify-you’)

More generally, non-sentential utterances—utterances lacking an overt predicate—are ubiquitous in conversations (see examples (1)–(3) above and elsewhere). (de Waijer, 2001) provides figures of 40%, 31%, and 30% respectively for the percentage of one word utterances in the speech exchanged between adults and infant, adult and toddler,
and among adults in a single Dutch speaking family consisting of 2 adults, 1 toddler and 1 baby across 2 months.

Non-sentential utterances are not a motley crowd. Recent studies have shown that they can be reliably classified into a small number of categories, revolving around the commonality in semantic resolution process (see e.g. (Fernandez & Ginzburg, 2002; Schlangen, 2003)). And yet, of course outside a conversational context a non-sentential utterance has little content—(7) illustrates that this same form can receive highly diverse contents from a wide range of sources: a previously uttered question, a question implicit in a particular domain, and as a correction

(7)  
   a. B: Four croissants.  
   b. (Context: A: What did you buy in the bakery?) Content: I bought four croissants in the bakery.  
   c. (Context: A smiles at B, who has become the next customer to be served at the bakery.) Content: I would like to buy four croissants.  
   d. (Context: A: Dad bought four crescents.) Content: You mean that Dad bought four croissants.

Thus, the competence in producing and understanding such utterances involves the context in an unavoidable way, including, as exemplified in (7b), how utterances fit in with social interaction. Conversely, matters of form can themselves, in the general case, require reference to the context. It was already pointed out by Lakoff (1971) and Morgan (1973)—though subsequently largely forgotten—that non-sentential utterances provide evidence that grammaticality cannot be adjudged context independently, i.e., simply by considering the morphosyntactic properties of a string. (8a,b) involve two virtually synonymous questions that lead to distinct contexts. (8a) is compatible with a possessive
NP as response, but not with a nominative NP, whereas in (8b) this pattern is reversed.

(8) a. A: Whose book did you borrow? B: Jo’s. /# Jo

   b. A: Who owns the book you borrowed? B: # Jo’s. / Jo. / It’s Jo’s.

Viewed from the perspective of the fragment, this pattern suggests that the fragment Jo’s has a presupposition that, to the extent its antecedent derives from a linguistic utterance, it must bear genitive case. Cross-turn dependencies of this kind are common among various types of non-sentential utterances, across a wide range of languages (Ross, 1969; Merchant, 2001; Ginzburg, 2012; Sag & Nykiel, 2011). What bears emphasizing is that such dependencies can stretch across many turns, particularly in multi-party dialogue, thereby reinforcing the need for this information to be in long-medium term representation of context: Ginzburg & Fernandez (2005) found that in the British National Corpus (BNC) over 44% of short answers have more than distance 1, and over 24% have distance 4 or more, as in the constructed example in (9):

(9)   A(1): Who is coming to the barbecue?

       B(2): the barbecue on Sunday?

       A(3): the 29th yes

       B(4): Sunday is the 28th.

       A(5): Oh right, yes the 28th.

       B(6): The one Sam’s organizing?

       A(7): Yes.

       B(8): Will it be on even if it snows?
A(9): Sam hasn’t said anything.

B(10): Right. Anyway, I’d guess Sue and Pat for sure, maybe Alex too.

### 3.2 References to the Interaction Situation

In this section we demonstrate the existence of data, much of it utterly ubiquitous, in which reference is made to the ongoing Interaction Situation, the events that take place in it, and their order.

#### 3.2.1 Clarification Requests

Plato was already at least implicitly aware of the fact that language enables one to explicitly address communicative aspects of an utterance: the Socratic dialogues are replete with examples of utterances whose primary function is to serve as clarification requests (CRs), in other words to indicate that some aspect of a prior utterance, typically its meaning, is unclear:

(10) a. Hermogenes: Yes; but what do you say of the other name?
    Socrates: Athene?
    Hermogenes: Yes.

b. Socrates: There is no difficulty in explaining the other appellation of Athene.
    Hermogenes: What other appellation?
    Socrates: We call her Pallas.

(From Cratylus, http://en.wikisource.org/wiki/Cratylus)

CRs can take many forms, as illustrated in Table 1, taken from (Purver, 2006), a taxonomy based on CRs occurring in the British National Corpus.

Table 1: A taxonomy for clarification requests (Table I in Purver 2006)
<table>
<thead>
<tr>
<th>Category name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context</td>
<td>A: Did Bo leave?</td>
</tr>
<tr>
<td>Wot</td>
<td>B: Eh? / What? / Pardon?</td>
</tr>
<tr>
<td>Explicit</td>
<td>B: Did you say ‘Bo’ / What do you mean ‘leave’?</td>
</tr>
<tr>
<td>Literal reprise</td>
<td>B: Did BO leave?</td>
</tr>
<tr>
<td></td>
<td>Did Bo LEAVE?</td>
</tr>
<tr>
<td>Wh-substituted Reprise</td>
<td>B: Did WHO leave?</td>
</tr>
<tr>
<td></td>
<td>Did Bo WHAT?</td>
</tr>
<tr>
<td>Reprise sluice</td>
<td>B: Who? / What? / Where?</td>
</tr>
<tr>
<td>Reprise Fragments</td>
<td>B: Bo? / Leave?</td>
</tr>
<tr>
<td>Gap</td>
<td>B: Did Bo ...?</td>
</tr>
<tr>
<td>Filler</td>
<td>A: Did Bo ... B: Win?</td>
</tr>
</tbody>
</table>

Providing explicit formal analyses of just about any of these classes is a formidable challenge for most existing formal grammatical frameworks. We highlight just several of the most significant issues.

The first point to note is that a number of these forms are ones whose sole analysis is as clarification requests—this applies to the classes Wh-substituted Reprise and to Gap. These constitute instances of forms the meanings of which cannot be analyzed in interaction-free grammar.

A second point relates to cross-turn parallelism. (Ginzburg & Cooper, 2004; Ginzburg, 2012) argue in detail that reprise fragments have two main classes of uses: one to request confirmation about the content of a previous sub-utterance; the other to find out about the intended content of a previous sub-utterance. Both uses have strong parallelism requirements: the former requires identity of syntactic category between source and target, as illustrated in (11a,b); the latter requires segmental identity between source and target, as exemplified in (11c). Parallelism of the latter kind seems needed also for the Gap class of CRs:

(11) a. A: Did she hit him? B: #he/ him  
     b. A: Was she biking? B: biking/cycling/#biked?
c. A: Did Bo leave? B: Bo?

(Intended content reading: Who are you referring to? or Who do you mean?)

Alternative reprise: B: Max? lacks intended content reading; can only mean: Are you referring to Max?)

A final point concerning the data of CRs involves anaphora: CRs typically involve reference to utterance tokens. This is, in fact, a more general requirement concerning quotative acts in dialogue, to which we return below.

(12) a. A: Max is leaving. B: leaving?

(=What does ‘leaving’ mean in the A’s sub-utterance, NOT in general.)

b. A: We’re fed up. B: Who is we?

(=Who is ‘we’ in the sub- utterance needing clarification)

3.2.2 Turn Taking

As first pointed out in the seminal paper (Sacks, Schegloff, & Jefferson, 1974), interlocutors manage turn allocation remarkably well. This has often been summarised as no-gap-no-overlap, though Heldner & Edlund (2010), based on a study of corpora in Dutch, Swedish, and English, conclude that sizeable departures from no-gap-no-overlap occur frequently, while cases with neither gap nor overlap are very rare: gaps with a duration above the threshold for detection of silences represent more than 40% of all between-speaker intervals in their material, whereas overlaps represent about 40% of all between-speaker intervals. Heldner and Edlund’s data suggests that a substantial share of (though clearly by no means all) speaker changes involve gaps long enough for the next speaker to react to potential signals occurring in the immediate vicinity of the speaker change. One grammaticized means for turn allocation are call utterances illustrated in (13) where the selected next speaker is addressed by name, though this can also be
affected by gaze or left partially underspecified:

(13) Sara: Ben you want some ( )? (# I’m not asking you Ben.) Ben: Well allright I’ll have a, (pause)
Sara: Bill you want some? Bill: No

(Example from footnote 13 of (Sacks et al., 1974))

Representing such a mechanism requires means of stating within the grammar information such as ‘referent of this NP is hereby offered the next turn’.

3.2.3 Order-dependent expressions

‘Metalinguistic’ expressions are expressions whose interpretation depends on the way other utterances have been pronounced, or on the order in which other expressions have been uttered. We will concentrate here on metalinguistic expressions whose interpretation is affected by the order in which other expressions are uttered or occur in a text, such as the former / the latter, vice versa, respectively, and the following (McCawley, 1970; Kay, 1989; Corblin, 1999; Yamauchi, 2006). The uses of these expressions we are interested in are illustrated in (14a) and (14b).

(14) a. Bob and John were at the meeting. The former brought his wife with him.
   (Quirk/Greenbaum)
   b. I think actors can teach dancers a lot, and vice versa. (From the British National Corpus.)

Former in (14a) has a different meaning from the meaning it has in expressions like George Bush, the former president of the US. This meaning can be specified informally as follows: the definite description denotes that element of a familiar set of individuals that is denoted by the first NP used to introduce an element of that set. The identifying property is “metalinguistic”: it refers to the order of elements in the text.

Vice versa in (14b) denotes the proposition obtained by exchanging the two arguments of
How can we make this informal semantics of order-dependent expressions more precise? One might think that the semantics of the former, at least, could be specified within a dynamic logic like Heim’s File Change Semantics (Heim, 1982), by assuming that an ordering exists on the set of file cards posited to underlie reference resolution. More specifically, one could propose that the sense of former under discussion is a predicate that is satisfied by the element of a set iff the file card associated with that element precedes the file cards associated with the other elements of the set. And indeed, a proposal along these lines was advanced by Corblin & Laborde (2001). Corblin and Laborde propose that the common ground consists of two parts: a part containing information about the propositional content of utterances, and a part so-called mentionelle which contains instead information about the mentions of file cards. Two points can be made about this proposal. First, the information mentionelle is in effect information about a subset of the utterances in the Interaction Situation—namely, the utterances of NPs. Second, in order to account for the entire range of order-dependent expressions, more is needed. This is because vice versa, in particular, can refer to the order of virtually any sentence constituents, not just noun phrases. In the famous Dorothy Parker joke I’m too fucking busy, and vice versa, for example, the constituents that get ‘switched’ are an adjective and an intensifier, and the switching affects their syntactic interpretation as well as their meaning.

3.2.4 Anaphoric references to events in the Interaction Situation

The class of antecedents for anaphoric expressions includes elements which can only be described by making reference to events in the Interaction Situation. Most current theories of discourse structure—e.g., SDRT (Asher & Lascarides, 2003)–assume that connectives involve some sort of implicit reference to illocutionary events. And indeed explicit references to illocutionary acts are also possible, as in the following example,

dancers can teach actors a lot (Culicover & Jackendoff, 2012).
where demonstrative *that* in the second utterance is a reference to the promise.

(15)  
  a. A: John, I promise I will help you with your homework.  
  b. B: That was a silly thing to say, as you won’t have any time.

Locutionary events can be referred to, as well. Webber (1991), for instance, pointed out to examples like (16), in which demonstrative *that* and pronoun *it* refer to the locutionary act performed with the first utterance.

(16)  
  a. A: The combination is 1-2-3-4.  
  b. B: Could you repeat that? I didn’t hear *it*.

### 3.3 Online Self-Repair/own communication management

As we saw in examples (1)–(3), conversations are littered with disfluencies—or, as we would prefer to put it, conversationalists continually utilize **own communication management** (OCM) devices to correct or modify their utterances or to gain extra time when facing lexical access or utterance planning problems. Although own communication management is viewed as a performance phenomenon in most formal grammatical treatments—a view explicitly rejected by psycholinguists (see, e.g., (Levelt, 1983; Clark & FoxTree, 2002; Ferreira, 2005)), the unity it displays with Other Communication Management (clarification questions and other–corrections) was noted already in the seminal paper (Schegloff et al., 1977). Probably the main substantive reason for pushing OCMs to the performance wastebasket is the assumption that they constitute noise. But in fact, far from constituting meaningless noise, OCMs participate in semantic and pragmatic processes such as anaphora, conversational implicature, and discourse particles, as illustrated in (17–19). In (17), the semantic process is dependent on the **reparandum** (the phrase to be repaired) as the antecedent:

(17)  
  a. Peter was, well, he was fired. (Example from (Heeman & Allen, 1999); the
anaphor refers to material in reparandum

b. A: Because I, any, anyone, any friend, anyone, I give my number to is welcome to call me (Example from the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992, ); implicature based on contrast between repair and reparandum: ‘It’s not just her friends that are welcome to call her when A gives them her number’)

Moreover, utterances containing disfluencies constitute a subclass of the antecedents for various linguistic expressions, including no and or:

(18)  

a. From yellow down to brown - no - that’s red.

b. Over the gree-, no I’m wrong, left of the green disk...

c. A: Is Bill coming? B: No, Mary is.

d. [A opens freezer to discover smashed beer bottle]
   A: No! (‘I do not want this (the beer bottle smashing) to happen’)

(19)  

a. We go straight on, or– we enter via red, then go straight on to green.  
( From Levelt (1989))

b. The design of or– the point of putting two sensors on each side. (From (Besser & Alexandersson, 2007)).

Non-disfluent speech is analogous to frictionless motion. Some of the time it is useful to ignore the effects of friction, but the theory of motion is required to explicate the existence and quantitative effects of friction. Whereas it seems plausible that not all disfluencies are consciously produced by the speaker, for the addressee they always form part of the string of phonemes perceived which needs to be parsed and interpreted.

Moreover, OCMs illustrate the primacy—or at least equal footing—of the speech event over grammatical form: as Levelt (1983) has observed, speakers will stop in ‘mid word’ when detecting error, as exemplified in (20a,b)—in the latter apparently the speaker replaces the beginning of the verb *instruct* with the less specific verb *do*; moreover,
speakers can stop in mid-utterance if the intended meaning seems to have been communicated—in (20c) D produces a clause headed by the complementizer *whether*, omitting a subordinating predicate (e.g., *is unclear/unlikely* etc) and both he and A laugh together about the mutually communicated content:

(20)

a. We can go straight on to the ye–, to the orange node. (Levelt, 1983)

b. Bee: y’know they(d) they do b- t! hhhh they try even harder than a- y’know a regular instructor. /Ava: Right. / Bee: hhhh to uh insr yknw do the class and everything. (example (18) from (Fox, Maschler, & Uhmann, 2010))

c. D: lots of secretaries do that, but it’s such a waste of time, but on the other hand you do meet / A : yes / D: secretaries. Whether you want to meet secretaries . . . A,D: (laugh) (From the London Lund corpus, (Svartvik & Quirk, 1980).)

Indeed, OCM utterances display an important characteristic of grammatical processes, namely *cross-linguistic variation*. This has been documented in some detail in comparative work between morphosyntactic aspects of repair on a wide range of languages by Fox and collaborators (e.g., (Fox et al., 1996; Wouk et al., 2009; Fox et al., 2010)); for phonetic analysis of cross-linguistic variation see (Candea et al., 2005), who compare fillers in Arabic, Chinese, French, German, Italian, European Portuguese, American English and Latin American Spanish. These studies demonstrate that language-specific features can be observed in the segmental structure of the fillers. French, for example, prefers a vocalic segment as filler realization, whereas English prefers vowels followed occasionally by a nasal coda consonant [m]. Moreover, while for some languages the vocalic support of the fillers might be a segment exterior to the vocalic system of the language, in all the eight languages the fillers’ vocalic support involves at least one of the vowels of their vocalic system.
There is some variation in how hesitation is typically expressed in various languages, as exemplified in (21). Indeed, some languages, e.g., Chinese exemplified in (21c), use demonstratives for this role:

(21)  
  a. ‘uh’ ‘um’ (English) (Clark & FoxTree, 2002, )  
  b. ‘euh’ ... (French): tu sais c’était un peu euh l’ambiance santa-Barbar-euh  
     ((De Fornel & Marandin, 1996), example (1a))  
  c. Chinese: ‘en’, ‘nage’ (literally ‘that’), ‘zhege’ (literally ‘this’) (Zhao & Jurafsky, 2005, )

Clark & FoxTree (2002), following an earlier proposal by (James, 1972) and based on data from the London Lund corpus, claim that the choice of *um* versus *uh* reflects an explicit choice by the speaker—the former being selected when the speaker faces a relatively significant difficulty which will lead to a longer wait for the resumption of the utterance (but see e.g., (O’Connell & Kowal, 2005; Corley & Stewart, 2008) for a dissenting view).

Recently, Tian et al. (2015) demonstrate significant preference for *um* v. *uh* among speakers of British English before self addressed questions (e.g., *What do they call it? What’s her name?*)—a clear signal of major difficulty, but no significant difference among speakers of American English (data from Switchboard); they also demonstrate marked preference for certain hesitation markers in Japanese and Chinese on the basis of distinct syntactic contexts. Wieling et al. (2015) demonstrate significant differences in the choice of *um* versus *uh* (and their cross-linguistic variants) both between male/female and younger/older speakers in four Germanic languages (Dutch, English, German, and Norwegian). This emergent body of work supports the claim that hesitation markers are words the choice between reflects explicit speaker intention.

Additional reasoning supporting the need for incorporating disfluency markers in the grammar are the following considerations: a child acquiring English needs to discover
that *no* can be used in a self-correction, but, for instance the closely related word *nope* cannot. A trilingual acquiring English, German, and French will need to learn that French *enfin* can be used in a self-correction, whereas English *finally* and German *schließlich*, which are often interchangeable with *enfin*, cannot be so used:

(22) Quand ma belle mère enfin quand ma femme apelle (De Fornel & Marandin, 1996, example (2a))

Conversely, Ginzburg, Fernandez, & Schlangen (2014) suggest that OCMs are also involved in grammatical universals. Based on evidence from 7 languages, they postulate the following:

(23) if NEG is a language’s word that can be used as a negation and in cross-turn correction, then NEG can be used as an editing phrase in backward-looking disfluencies.

### 3.4 Why there cannot be a global grammar: evidence from quotation

The phenomenon of direct quotation perhaps epitomizes the point of the paper: it is ubiquitous, it is subject to grammatical constraints, but features in few formal grammars (for some recent formal treatments see (Geurts & Maier, 2005; Potts, 2007; Bonami & Godard, 2008), but these do not form part of a large scale grammar). In a way this is not surprising since, as argued by Ginzburg & Cooper (2014), quotation is a challenge for any grammar G: for any string e deemed ungrammatical by G, one can produce via quotation a well formed string that includes e, hence undermining G. Thus, we can quote something that is ungrammatical in our own language as in (24a) or something that is in a different language to the one we are speaking (24b), sounds made by inanimate objects (24c), or the thoughts of non-humans (24d).

(24) a. Damien, who’s only four years old, said `I go’ed to Grandma’s’
b. Pelle, whose native language is Swedish, said `Jag har varit hos mormor’ (meaning “I’ve been at Grandma’s”)

c. The blender went `plplplpl’

d. [An article about an orphaned walrus arriving in a new zoo:] During [the orphan walrus’s] first look at a walrus, he was like, ‘What’s that?’ (New York Times, 22/01/2014)

Given the diversity of quotable stuff, one might very well think it beyond the remit of somebody writing a formal grammar of English to characterize everything that can occur between quotation marks in sentences like those in (24). Such a strategy is, however, not tenable, for reasons mostly pointed out already by Partee (1973), who provides a variety of examples where the form or the content of the quotation is referred to from outside the quotation as in (25).

(25)  a. ‘I talk better English than the both of youse!’ shouted Charles, thereby convincing me that he didn’t.

b. What he actually said was, ‘It’s clear that you’ve given this problem a great deal of thought,’ but he meant quite the opposite.

And indeed there is a fait bit of evidence that quotation is subject to general grammatical principles governing word order, ability to be embedded and psuedo-clefted, and semantic selection (Postal, 2004; Bonami & Godard, 2008). In particular, there are words in various languages that require direct quotation as their complements. In English the marker like and the verb go have a certain usage which requires a direct quotation as in (26a) and (26b) and does not allow an indirect quotation, as exemplified in (26c) and (26d). Such constructions exist in many, if not all, languages although they tend to be restricted to an informal spoken register, see e.g., French faire, genre, German quasi, Italian tipo and Swedish typ/ba.
Moreover, all natural languages seem to have direct quotation of some kind. Children use direct quotation from their earliest utterances (Ginzburg & Moradlou, 2013). Given the ubiquity of quotation in natural language, linguists need to explicate the mechanisms it employs. Indeed, one is obligated to do so in a way that offers an answer to the question: why, rather than being a heterodox linguistic process, is in fact quotation so straightforward? We will suggest one such answer below. Whatever one proposes, it seems clear that direct quotation is a grammatical construction where reference is made to an interaction act, constrained via a similarity relation that needs to hold between the quoted material and the original act—Ginzburg & Cooper (2014) argue that the nature of the similarity relation is a contextual parameter of this construction. Moreover, it forces the grammar to be an intrinsically open system (Harris, 1979; Postal, 2004).
3.5 Pointing, gestures and the Interaction Situation

The view of the role of pointing and other gestures in communication as discussed in Section 2, that essentially abstracts away from the Interaction Situation, has been challenged in a number of ways.

**Pointing** Extensive empirical work by Lücking, Pfeiffer, Rieser and colleagues (Kranstedt et al 2006; Lücking, Pfeiffer, & Rieser, 2015) using highly sophisticated recording and visualization equipment has demonstrated that pointing gestures seldom if ever function as unique identifiers of a demonstratum, as proposed Kaplan (1978). In all but the simplest situations, the identification of the demonstratum among the objects in the pointing cone identified by a pointing gesture is a complex reasoning process involving consideration of a number of additional aspects of the Interaction Situation.

Beyond this, Clark (2003) showed that pointing is neither the only nor the prototypical way to carry out a demonstration. For instance, a customer can felicitously demonstrate to the teller in a supermarker the referents of a demonstrative like *These two things over here* by placing the two objects on the counter rather than merely pointing at them.

**Interactional role of other gestures** Kendon (2004) distinguishes between two types of gestures: gestures that contribute to what he calls the ‘referential meaning’ of the utterance (discussed in his Chapters 9 to 11) and ‘pragmatic’ gestures (discussed in Chapters 12 and 13). Among the latter, there are several whose function is to manage aspects of the Interaction Situation. These include gestures whose function is to indicate to whom a current utterance is addressed, and several gestures that play a role in turn-taking: for instance, indicating that the speaker is holding the floor, or raising a hand to request a turn, or pointing to indicate the next to hold the floor.

Further evidence is provided by examples like (27a), that exemplifies a wordless exchange mediated solely by display and gesture, which corresponds to a
question/answer pair, as in either (27b), where the question is implicit and the answer is a non-sentential utterance, or (27c), where the question is explicit and the answer is a non-sentential utterance. This indicates the need for a mechanism that unifies all three cases, given the intuitive synonymy.

(27)  
   a.  Owner: (displays three fresh fish on a platter) Clark: (points at one of them)  
      (From (Clark, 2012): example (32))  
      b.  Owner: (displays three fresh fish on a platter) Clark: That one.  

Note that whereas meaningful utterances of just about any modality can be clarified using a construction such as *what do you mean* (28a,b), clarification via repetition is limited to speech (28c,d). The first fact indicates that such utterances are viewed as having a content on a par with linguistic speech; the second fact demonstrates that clarification via repetition is not ‘anything goes’ and involves a construction that has clearly defined selectional restrictions (in contrast with certain other quotative constructions discussed in section 3.4):  

(28)  
   a.  A : (laughs). B: What do you mean he he? / Why are you laughing?  
   b.  P1: my spine’s like (*gestures spine shape*) (1.0 sec) like that (0.9 sec) P2:  
      like this? (*gestures to clarify spine shape, quizzical face*) (From the corpus  
      described in (Healey et al., 2015))  
   c.  A : (laughs). B: (laughs) [cannot mean ‘what’s the meaning of your  
      laughter?’]  
   d.  A: (makes strange gesture) B: (repeats A’s gesture) [cannot mean ‘what’s  
      the meaning of this gesture you just made?’]

Finally, we note two interactions between phenomena described earlier. First, the possibility of quoting gesture, as in (29):
(29) Claire: How pleasant, mum’s being sick everywhere.

    I said erm is there a problem? (laugh) (vomiting noise) No. Not a problem. (BNC, KPH)

Second, the finding of Cook, Jaeger, & Tanenhaus (2009) that when speakers produce structures that they themselves usually disprefer, they are more likely to produce OCM utterances and to produce gestures.
Evidence from other disciplines

4.1 Language acquisition

That conversational interaction has a crucial role to play in language acquisition is at this point not very controversial (Hoff, 2006, Pruden et al, 2006): passive exposure to language, as afforded by exclusive input from television (e.g., Kuhl, Tsao, & Liu, 2003) is insufficient to acquire native competence. Indeed, children often do not acquire native-like competence in the absence of peers as a source of native input (Oller and Eilers 2002). The converse case, interaction with a language model lacking, is exemplified by deaf children born to non-signing hearing parents. In this case there is invention by the child of a communicative system, but one that fails to develop into fully articulated language (Goldin-Meadow, 2003). Quantity and type of interaction play a significant role: there is extensive evidence about differences in amount and type of utterances children are exposed to across distinct social socioeconomic status (SES), most famously (Hart & Risley, 1995) reported a ratio of approximately 4 : 2 : 1 for the total words heard by, respectively, American children of high SES parents middle SES parents, and lower SES parents. This is strongly correlated with speed of acquisition: by 3 years of age, the mean cumulative recorded vocabulary for the higher SES children was over 1000 words and for the lower SES children it was somewhat less than 500, whereas other studies show similar large effects on grammatical development (e.g., Snow, 1999).

It has also been known for some time that interactional abilities such as turn-taking and maintaining joint focus of attention are critical precursors for language development (Scaife & Bruner, 1975; Carpenter, Nagell, & Tomasello, 1998, Dominey & Dodane, 2004). Moreover, recent work has shown that pointing is specifically implicated in the acquisition of the lexicon and simple syntactic structures (Tomasello et al, 2007, Rowe & Goldin-Meadow, 2009).
A bioecological approach to development like that championed by Bronfenbrenner (Bronfenbrenner and Morris, 2006) offers tools for integrating different sources of environmental effect in language acquisition (Hoff, 2006), given its emphasis on the importance of interaction but also on a quantitative approach to heritability (Bronfenbrenner and Ceci, 1994). An important question is whether this can be integrated with specifically linguistic theories of acquisition to identify the mechanisms through which genotypes are transformed into phenotypes’ (Bronfenbrenner and Ceci, 2006).

The two main approaches to the acquisition of grammar at present are nativism, inspired by Chomskyan assumptions (Chomsky, 1965; Snyder, 2007) and the usage–based approach (Tomasello, 2003). These two approaches differ radically on a number of dimensions, most particularly with respect to the role of learning; for nativism this is limited to words and how these relate to Universal Grammar, whereas the usage-based approach highlights the importance of domain-general learning mechanisms such as analogy, entrenchment, and automatization.

As things stand, however, neither nativist, nor the usage-based approach has advanced an explicit theory that would enable one to make clear predictions about how the grammatical system of a child evolves at various points as a result of conversational interaction with her parents. This is, in part, because, with very few exceptions (Ginzburg & Moradlou, 2013; Jackendoff & Wittenberg, 2014), the early stages of linguistic competence have not been formally described, presumably because of the significant challenge they pose for existing grammar frameworks. Most crucially, because neither approach has as yet developed an explicit theory of learning in spontaneous conversation, which will inter alia incorporate the positive effects arising from the provision of non-negligible amounts of ‘negative evidence’ by parents (Saxton, 2000; Chouinard & Clark, 2003) using other communication management. We believe that progress on this crucial task requires an approach to grammar like the one outlined here.
4.2 Cognitive neuroscience

Earlier claims regarding e.g., the role of Broca’s area in the processing of transformations (see, e.g., (Bambini, 2012) for a general survey of the neuroscience of language, and (Grodzinsky, 2003) for an hypothesis about transformations in the brain) notwithstanding, there is still a substantial disconnect between the research programs of cognitive neuroscience and theoretical linguistics, and the hypotheses that get formulated in those camps (Poeppel & Embick, 2005; Grimaldi, 2012). The primary interest of cognitive neuroscientists studying language, neurolinguists, is identifying the areas involved in different aspects of language processing; and there is now converging evidence that several areas are involved, above all the frontal lobe (e.g. Brodmann areas 44–Broca’s area–45, 46, and 47), the temporal lobe (e.g., the superior temporal lobe, STL - Wernicke’s area - and the superior temporal gyrus, STG), and parietal lobe (e.g., the angular gyrus) (Bambini, 2012; Grimaldi, 2012). Such evidence does not clearly support either the claim of a separate ‘faculty of language’, or the existence of a division between competence and performance (Grimaldi, 2012).

Some of the aspects of language use that we are proposing are governed by grammar, in particular turn-taking, have been studied in the field of neuropragmatics (Van Berkum, 2010; Bambini, 2012) but such studies show that the areas involved in such processing are the same, or very closely related, to those involved in aspects of language interpretation more traditionally accepted as involving competence. In fact, such studies tend to show that involvement in those aspects of language use results in greater activation of some of the areas associated with language processing. For instance, (Jiang et al 2012) found, using fNIRS, that face-to- face dialogue results in increased activation in the inferior frontal cortex in comparison with back-to-back communication, or back-to-back monologue. And the comparison with face-to-face monologue strongly suggests that the difference in activation is primarily based on turn taking and body
Evidence concerning the timing of these interpretive processes doesn’t clearly support their isolation from conventional aspects of grammatical interpretation, either. Evidence by, e.g., (Egonova, Pulvermueller, & Shtyrov, 2014) suggests that speech act identification and interpretation takes place rapidly—in fact, more rapidly than certain types of lexical-semantic processing.
5 Interaction Grammar: principles and illustration

Historically, one of the reasons for the relative neglect by linguists of phenomena such as those discussed in Section 3 has undoubtedly been the lack of adequate formal tools to describe their grammar. One of the key contentions of this paper, however, is that this is no longer the case, and that several grammatical frameworks now exist which provide the tools to characterize the grammar of such phenomena.

In this section we will first identify the fundamental assumptions that in our view a linguistic framework needs to share in order to constitute an instance of “Interaction Grammar”; we then sketch informally how such assumptions can provide an account for the variety of phenomena discussed in Section 3. We should emphasize that there are various ideas currently being proposed about how Interaction Grammar should be achieved. We are therefore not presenting a single theory, but rather making a claim about what the guiding principles and the key components of such a theory would be. All the components we posit are used in the accounts we sketch below, hence discarding any component entails demonstrating its explanatory dispensability.

5.1 Key theoretical assumptions

The interactionist view of grammar involves making at least the following assumptions.

1. **Interaction Situation reference** Grammar makes essential reference to a dynamically updated Interaction Situation which indicates what is happening as the interaction takes place, along with some record of what has happened already.

2. **Sign instantiation**
Grammar makes essential reference to certain audio-visual-gestural events that occur in the Interaction Situation: uttering sounds, pointing, gesturing, etc. Furthermore, in grammar

(a) **Incremental classification**: such events are classified into grammar-relevant types (signs) in incremental fashion by conversational participants.

(b) **Partiality**: The classification process can be partial, where the type does not uniquely classify the event, thereby triggering repair.

(c) **Non-monotonicity**: The classification process can be non-monotonic: the type assigned to an event can change as a consequence of repair.

3. **Event types in grammar rules** Linguistic generalizations and procedures are expressed not solely in terms of the events themselves but also in terms of types of events (or situations).

4. **Event type inference** Event types are used in rules which specify the enrichment of the Interaction Situation by propositional and erotetic inference.  

5. **Language in flux** The class of grammatical types can be modified during interaction.

Assumption 1, *Interaction Situation reference*, is relatively uncontroversial: any grammar that treats indexicals like *I*, *you*, and *now* needs to effect reference to the Interaction Situation somehow. However, the orthodox treatment (following (Kaplan, 1978)) is for this reference to be viewed as external to the grammar, formulated as indices relativising the evaluation of sentences; the extent of indexicality we propose and its explicitness yields significant novelty. *Language in flux*, by contrast, is not assumed in any major approach, but as shown in Sections 4 and 5, it is a key assumption for both language acquisition and communication management (repair). But the key innovations in
Interaction Grammar are *Event types in grammar rules, Event type inference*, and *Sign instantiation*. The latter involves several sub-assumptions, which are pairwise independent (so a grammatical framework might satisfy one without satisfying one of the others.). As we will see, these assumptions have several controversial consequences for a more traditional view of grammar.

For concreteness we will assume a specific characterization of the notion of Interaction Situation, developed in the dialogue semantic framework KoS (Ginzburg, 2012); though there are a variety of alternative theories of this notion, from the original formulation of discourse situation in (Barwise & Perry, 1983) to PTT (Poesio & Rieser, 2010, 2011), these theories share with KoS many of the key hypotheses about the Interaction Situation.

In KoS, the **dialogue gameboard**, inspired by Lewis (Lewis, 1979), represents information that arises from publicized interactions. Its structure is shown in (30). The Spkr, Addr fields track turn ownership; Facts represents conversationally shared assumptions; VisualSit represents the dialogue participant’s (view of) the visual situation and attended entities; Pending and Moves represent respectively moves that are in the process of/have been grounded; QUD tracks the questions currently under discussion, though not simply questions *qua* semantic objects, but pairs of entities which we call **InfoStrucs**: a question and an antecedent sub-utterance. This latter entity provides a partial specification of the focal (sub)utterance, and hence it is dubbed the **focus establishing constituent** (FEC) (cf. parallel element in higher order unification–based approaches to ellipsis resolution e.g. (Gardent & Kohlhase, 1997); (Valduví, 2015) relates the FEC with a notion needed to capture *contrast*).
We will use this basic characterization of the Interaction Situation to present bare-bones analyses of a selection of the phenomena discussed in Section 3. Our discussion here is informal, but in virtually all cases formally worked out treatments already exist to which references are provided.

5.2 Sign instantiation and its consequences

One of the types of events that are recorded in the Interaction Situation according to the sign instantiation hypothesis are utterances. Specifically, we assume that as the result of utterances taking place, the Interaction Situation is dynamically updated by recording pairs

\(<\text{utterance event}, \text{utterance type}>\)

where an utterance type is the Interaction Grammar equivalent of a sign in the sense of sign-based grammars such as Head Driven Phrase Structure Grammar (HPSG, (Pollard & Sag, 1994; Ginzburg & Sag, 2000; Sag et al., 2003)), Categorial Grammar (see e.g. (Calder, Klein, & Zeevat, 1988; Moortgat, 1997)), or in versions of Lexical Functional Grammar (see e.g. (Muskens, 2001)).
A pair \( \langle u, T_u \rangle \) indicating the occurrence of an utterance event \( u \) of type \( T_u \) is called a **locutionary proposition**. For instance, suppose that \( A \) utters (31a). Then the **Pending** constituent of the Interaction Situation is augmented with the locutionary proposition in (31b), stating the occurrence of utterance event \( u_q \) of type \( \text{Say}(A, \text{‘Did Bo kowtow?’}) \).

(31) a. \( A: \text{Did Bo kowtow?} \)
   
   b. \( \langle u_q, \text{Say}(A, \text{‘Did Bo kowtow?’}) \rangle \)

In fact, in versions of Interaction Grammar like KOS or PTT every sub-utterance of \( u_q \) expressing a constituent of \( u_q \) gets recorded as a separate locutionary proposition—e.g., the utterance event \( u_{kowtow} \) of uttering the word *kowtow*:

(32) a. \( A: \text{Did Bo kowtow?} \)
   
   b. \( \langle u_{Did}, \text{Say}(A, \text{‘Did’}) \rangle \)
   
   c. \( \langle u_{Bo}, \text{Say}(A, \text{‘Bo’}) \rangle \)
   
   d. \( \langle u_{kowtow}, \text{Say}(A, \text{‘kowtow’}) \rangle \)

In this paper, we assume two main types of verbal interaction events—Say and Ask—as well a few other non-verbal interaction events discussed below.

### 5.2.1 Other Communication Management

As discussed in Section 3.3, the grammar makes available various constructions whose primary function is to request clarification about prior utterances. We discuss here two cases; for detailed analysis see (Ginzburg & Sag, 2000; Purver, 2006; Ginzburg, 2012). An analysis of **sentential reprises** such as (33) involves a construction which, via reference to the Interaction Situation, builds a content in the following way: the maximally pending utterance serves as the proposition from which a question is formed, indicated here using the notation \(?p\)—zero or more argument roles are queried, corresponding to referential elements that cannot be resolved in context:\(^{11}\)
For an utterance like (34a), we need to say more about the reasoning an interlocutor makes when posing a clarification question. We assume that after every utterance the addressee engages in monitoring the incoming utterance $u0$: if she thinks she understands it—she can classify $u0$ with a fully instantiated sign, she responds accordingly; if not, taking as input her partially instantiated locutionary proposition, she has a right to accommodate into the context one of a small number of questions concerning any sub-utterance of $u0$ (Ginzburg & Cooper, 2004; Purver, 2006; Ginzburg, 2012). Thus, for any sub-utterance $u1$, the grammar enables reference to the question ‘what did prev-spkr mean by $u1$’ constrained by segmental phonological parallelism with $u1$. In other words, we assume the existence of a construction whereby a phrase identical to a sub-utterance $u1$ of the previous utterance can express a question like (34b), where we use the function $\rightarrow$ to map an utterance event $u$ to its content $x$ (Poesio & Rieser, 2010, 2011):

(34) a. A: Did Bo kowtow? B: kowtow?
   b. $?x(u_{kowtow} \rightarrow x)$ (“what did A mean by uttering ‘kowtow’?”)

What price do we need to pay to develop an account like this one of (34a)? The main cost involves the context: via Sign instantiation and Event type inference we assume that interlocutors maintain highly structured representations of utterances to enable them to engage in clarification question accommodation. Specifically, representations which specify the morphosyntactic and meaning representation for each sub-utterance, given the fact that each sub-utterance down to the level of the word is potentially clarifiable (Poesio, 1995; Poesio & Muskens, 1997; Purver, Ginzburg, & Healey, 2003, (forthcoming); Poesio & Rieser, 2010, 2011).
As far as the grammar goes, the cost is this: the ability to specify constructions which make reference to elements of QUD. This latter requirement is currently supported by much evidence (Ginzburg, 1994, 2012; Roberts, 1996, 2004).

5.2.2 Quotation

Given the ubiquity of quotation in natural language, linguists need to explicate the mechanisms it employs. Indeed, as we suggested earlier, one is obligated to do so in a way that offers an answer to the question: why, rather than being a heterodox linguistic process, is in fact quotation so straightforward? The short answer, we suggest, is that this is because quotation involves entities and mechanisms utilized ubiquitously during dialogue processing: in particular, Sign Instantiation.

How does this apply to quotation? (Ginzburg & Cooper, 2014) postulate that pure quotations denote signs and direct quotations denote locutionary propositions. We illustrate how this applies to direct quotation briefly.

Direct quotation involves providing a demonstration of a previous communicative act $u$ (or in extreme cases a sound or gesture act imbued with communicative intent.) (de Cornulier, 1978; Clark & Gerrig, 1990). What varies with context is how similar the demonstration is going to be to $u$ (does the demonstration use the same language? does it filter away disfluencies? how close in terms of content is it to $u$?). By representing a direct quotation in terms of $u$ (the original act) and $T$ (the type corresponding to the demonstration), we can specify the similarity required in context.

A predicate embedding a direct quotation like English like, go or French faire is then posited to select for a locutionary proposition $<u,T>$ and to predicate of the content of $u$.

Thus, in (35a), A makes an utterance in French including the hesitation marker euh. B reports this utterance in English using the utterance No way I’ll do it which has filtered away the hesitation and whose content B views to be sufficiently similar to $u_A$, A’s utterance:
(35) a. A: Je euh le ferai genre absolument pas.
   b. content(uA) = Not(Do(A,d))
   c. B: He went ‘like no way I’ll do it’.
   d. B’s direct quotation of A: Say(B, Say(A, (uAT^no way I'll do it))))
5.2.3 Own Communication Management

Dealing with OCM does not require much change as far as context goes: the monitoring and update/clarification cycle is modified to happen at the end of each word utterance event—or in principle more frequently (Brennan & Schober, 2001)—, and in case of the need for change, a clarification question gets accommodated into QUD. Overt examples for such accommodation is exemplified in (36):

(36) a. Carol: Well it’s (pause) it’s (pause) er (pause) what’s his name? Bernard Matthews’ turkey roast. (BNC, KBJ)
   b. A: Here we are in this place, what’s its name? Australia.

The answer to this question is then used as the alteration and this triggers an update of the representation of the utterance (Ginzburg et al., 2014).

While the contextual background to OCM requires little change to the view of context outlined previously, accounting for OCM requires considerable changes in the outlook of the grammar. Specifically, it requires:

1. an incremental and non-monotonic view of utterance construction.
2. ‘non-grammatical’ speech events to be incorporated within the domain of the grammar.

This latter assumption is required since words and collocations that constitute ‘editing phrases’ (e.g., No, Or, I mean) select for utterance events which can contain ‘ungrammatical’ aspects.

Hence, the status of the grammar shifts radically, potentially in line with views that argue for intrinsic gradience (Lau, Clark, & Lappin, 2014). It now characterizes as ‘well formed’ speech events that contain ill formed parts, albeit ones that have been corrected, for instance the German and Hebrew ones in (37a,d); a native speaker can distinguish these from potential utterances such as (37b,c,e,f) with no corrections or where the correction has gone awry:(37) a. der der die Batterie die versorgt (example (20), (Fox et al., 2010))
b. der Batterie die versorgt

c. die die der Batterie die versorgt

d. kaxa she hi amda he’emida oto leyad haluax. (example (26), (Fox et al., 2010))
e. kaxa she hi he’emida oto leyad haluax.
f. kaxa she hi he’emida amda oto leyad haluax.

5.2.5 Interjections and turn assignment

Consider a word like marḥabteyn. As we discussed in section 3.1, this word is used as a response greeting by Bilal just in case the initial greeting by Awda was marḥaba.

In a grammar which enables reference to the Interaction Situation, this is easy to capture: such a word has a presupposition about the form and content of the previously grounded move, that its form was marḥaba and content a greeting.

What of turn assignment utterances, as in (13)? As with greetings, in a grammar that allows reference to the Interaction Situation, which tracks turn holders, an utterance which expresses a wish about a projected turn holder is easy to encode.

5.3 Non-sentential utterances

In section 3.1, we pointed out that the content one assigns to a non-sentential utterance like four croissants can vary widely, with the sources for the different contents ranging from a previously uttered question through domain–specificity and to a correction. We have also emphasized that different non-sentential utterance constructions exhibit morphosyntactic and/or phonological parallelism with their antecedents, which in the case of short answers can be maintained across multiple turns. This means that not only the combinatorial semantics of non-sentential utterance constructions integrates information from the Dialogue Gameboard, but that this is also potentially true of the morphosyntactic and phonological specifications of such constructions. Such information
needs to be projected into the context, as we have already observed in the case of repair
constructions, maintained, in parallel with QUD–oriented information.

We claim that it is only with a theory of interaction that structures the context
appropriately that we can capture the uniformity underlying such utterances via a
construction type whose content field is specified as in (38e) and such that the fragment
maintains categorial parallelism to its antecedent in QUD: its predicate is the question
under discussion, its subject the bare fragment utterance, a generalization of a rule
proposed already in (Hausser & Zaefferer, 1979).

(38) Declarative-fragment-clause:

\[
\text{Cont} = \text{DGB.\text{MaxQU}(u\text{-frag} \rightarrow x)} : \text{Proposition}
\]

\[
\text{u-frag.cat} = \text{MaxQU.\text{fec.cat}} : \text{Syncat}
\]

For a detailed analysis of a wide range of NSU types found in the BNC see (Fernández,
2006; Ginzburg, 2012).

What of the cases such as (27)? The answers get introduced into the semantics via
mechanisms discussed in (Lücking et al., 2015), whereas the question via genre-based
inference (Larsson, 2002; Ginzburg, 2012).

5.4 Order-dependent expressions

One of the key theoretical assumptions of Interaction Grammar listed in Section 5.1, *Sign
instantiation*, specifies that the Interaction Situation includes a locutionary proposition
for every single word. Using this assumption we can provide an exhaustive account of
order-dependent expressions.

Using the notation introduced above \(<u, \Phi>\) to state that utterance event u is of type Φ,
and assuming again a function \(\rightarrow\) mapping utterance events to their content, we can say
that the result of uttering the NP *Bob* is to update the current utterance (the maximal element of Pending) by adding to it the two conditions in (39). The first one records the utterance $u$ by A of the word-string *Bob*; the second one records that the content of the utterance event $u$ is the object denoted by b. (We assume here a ‘natural’ semantics for proper names as proposed by Partee, together with type raising operations.) Subsequent utterances of the expressions *and, John*, etc. update the common ground in a similar fashion, by adding new utterance events preceded by $u$.

(39) $\{<u, \text{Say}(A, \text{‘Bob’})>, u \rightarrow b\}$

It should be easy to see how this view of the utterance events in the Interaction Situation can be used to specify the interpretation of order-dependent expressions like the former one and vice versa. We sketch here how such meanings can be characterized.

**the former one** The content of an event of uttering an NP of the form *the former N* is that element $x$ of a set $X$ of familiar objects which is also the content of the first utterance event $u_1$ among those used to introduce the elements of $X$.¹⁴

The required constraints on the Interaction Situation are imposed by assigning to *former* the following interpretation. Say that $u$ is an event of uttering *former*:

$<u, \text{Say}(A, \text{‘former’})>$.  

The interpretation of $u$ consists a ‘linguistic’ part (the content of $u$) and a ‘metalinguistic’ one. This second part imposes conditions on the Interaction Situation: namely, the requirement that two events of uttering nominals $u_1$ and $u_2$ occurred in the Interaction Situation, $u_1$ preceding $u_2$ and having content $x$. The linguistic part of the meaning of *former* then specifies that the content of the utterance $u$ of the adjective *former* is a predicate modifier specifying the restriction that the object to which the predicate is applied must be equal to $x$.  

vice versa the content of an event of uttering the word form vice versa conjoined with an utterance like (14b) with constituents $u_1 \ldots u_n$ is obtained by applying the usual rules of semantic interpretation to combine the contents of $u_1 \ldots u_n$, after having switched two contextually specified utterance events $u_i$ and $u_j$ (a proposal of this type was made, e.g., by Culicover & Jackendoff (2012)).

5.5 Anaphora

We have shown in, e.g., (Poesio & Rieser, 2010, 2011) that by adopting an interactionist approach to grammar the examples discussed in Section 3.2.4 can be analyzed within a treatment of anaphora that is a natural extension of Discourse Representation Theory (Kamp & Reyle, 1993) and is closely related to, e.g., the proposals in (Asher & Lascarides, 2003). In such theories, updating the Interaction Situation with new locutionary or illocutionary events makes new discourse referents available just like events in the situation under discussion. As a result, implicit anaphoric references such as those in (40a), repeated here for convenience, can be handled precisely as shown in (40b), i.e., references to events like $ce_1$ and $ce_2$ that are recorded in the Pending or Moves components of the Dialogue Gameboard:

(40) a. Although MSG has been blamed for a variety of symptoms, it has been vindicated

   by scientific research.

   b. $<ce_1, \text{assert(writer, 'MSG has been blamed for a variety of symptoms')}>$, $<ce_2, \text{assert(writer, "MSG has been vindicated by scientific research")}>$, concession($ce_1$,ce_2 )

Explicit references to illocutionary acts as in (41b) can then be handled as references to such events:

(41) a. A: John, I promise I will help you with your homework.
<ce1, promise(A,'''A will help John with his homework'')>

b. B: That was silly, as you won’t have any time.

<ce2, assert(B,'ce1 was silly as A won’t have any time')>

The references to locutionary events as in example (16b) from (Webber, 1991) (repeated as (42b) can be analyzed in a similar way, i.e., as references to utterance u1 of ‘the combination is 1-2-3-4’) provided we assume that not only illocutionary events, but locutionary events like u1 are recorded of the Interaction Situation:

(42) a. A: The combination is 1-2-3-4.
   b. <u1, Say(A,'the combination is 1-2-3-4')>
   c. B: Could you repeat that? I didn’t hear it.
   d. <u2, Say(B,'could you repeat u1? I didn’t hear u1')>

5.6 Pointing and Gestures

A grammatical framework in which grammar imposes constraints on the Interaction Situation is naturally suited to specify grammatical constraints on other aspects of communication such as gestures and pointing, as these are just other types of events whose occurrence is recorded in the Interaction Situation. In (Rieser & Poesio, 2009) we proposed that propositions of the form

<g,G(A)>

where G is a type of gesture, are recorded in the Interaction Situation as result of the performance of a grammatically relevant type of gesture by A. An example of grammatically relevant gesture is pointing:

<p,point(A)>

(from which we can indirectly infer, following the type of reasoning studied by Lücking et al, that <p, point-at(A,φ,)>). A multimodal grammar for the integration of pointing and speech based on this treatment of gestures in the Interaction Situation was proposed
in (Poesio & Rieser, 2009). Clearly, the framework could also be used to provide an account of gestures referring to other aspects of the Interaction Situation—e.g., for turn-taking.
6 Discussion

6.1 Moving the boundary between competence and performance

Without wanting to commit ourselves to a definite position concerning the existence of a competence / performance distinction, let us assume, initially, for argument’s sake that such a distinction is tenable as the basis for a theory of the human language faculty. What we have shown in this paper is that the boundary as commonly drawn is entirely artificial as it leaves out a host of key aspects of interaction that are clearly governed by ‘grammar’ under any sensible notion of what a ‘grammar’ is.

A secondary but still key aspect of our proposal is that this redrawing of the boundary does not necessarily involve abandoning any hope of providing a formal account of interaction. To be sure, there is still a lot of work to be done in developing a formal ‘Interaction Grammar’ framework that may provide as productive a foundation for theories of the extended notion of grammatical competence as the ‘standard theory’ that emerged in the 1970s and 1980s from the work of Chomsky, Montague, Partee, Bresnan, Sag, and many others. But we hope that for all its necessary sketchiness the proposal in Section 5 shows what the essential ingredients of such a formalism would be; much more detailed developments have appeared in, e.g., (Ginzburg, 2012).

This proposal also displays what we hold to be a key property of any future framework of this type: i.e., that it doesn’t overly muddy the grammatical baby with the interactional bathwater, i.e., that it is an extension and a generalization of the frameworks currently in use so that it does not require rethinking current grammar theory wholesale, as for many phenomena there already exist satisfactory accounts. Also, such an extension and generalization would allow linguists interested in phenomena that do not appear to involve reference to the Interaction Situation to use only the formal machinery that is strictly required.
A third contention we have tried to exemplify throughout is that redrawing the boundaries this way will make work on grammar by theoretical linguists much more relevant to sister disciplines such as computational linguistics, conversation analysis, corpus linguistics, psycholinguistics, speech processing, the study of multimodal interaction, or cognitive neuroscience that in recent years have had to develop their own foundational frameworks as the formal tools provided by theoretical linguistics were too limited (Ferreira, 2005; Poeppel & Embick, 2005; Steedman, 2013).

6.2 The grammar-pragmatics boundary

We expect several readers of this paper will react by saying ‘interesting phenomena, but this is not grammar, it’s pragmatics.’ Charting the semantics/pragmatics boundary is no easier than identifying the boundary between competence and performance (for some recent discussions, see (Recanati, 2010; Stojanovic, 2013; Lepore & Stone, 2014), and there are certainly influential proposals suggesting that pragmatics intrudes in various incontrovertibly grammatical processes (Levinson, 2000; Ariel, 2008). Avoiding these difficult issues here, we note that of the five classes of phenomena discussed, *Grammar across turns, Online repair, Genre dependent grammar, Gesture interactivity* are all concerned incontrovertibly with structural issues or issues of meaning composition. This leaves the class of phenomena concerned with reference to the Interaction Situation: but we pointed out that Other Communication Management constitutes the primary/literal meaning of a number of words and constructions, hence integrating these in grammar is as justified as integrating tense, which involves ordering relations between a described event and an utterance event (in our terminology—the *Interaction Situation*).

6.3 Rethinking competence v. performance as black box v. white box testing
The competence/performance distinction is prima facie attractive because it enables one to separate analysis of “the linguistic phenomena” from the specific details of how they get processed. The problem, we think, is that this reasonable desideratum has lead to a highly selective and misleading view of what are the ‘rule governed’ phenomena associated with language. We think a better construal of this separation could be drawn from computer science, which offers the distinction between black box and white box testing (Patton, 2006): the former pertaining to the functionality of an application without peering into its internal structures or workings, whereas the latter involves trying to assess functionality, in part, by examining the implemented code.
Acknowledgements

We thank Anne Abeillé, Gennaro Chierchia, Eve Clark, Herb Clark, Rebecca Clift, Robin Cooper, Nadine Glas, Mirko Grimaldi, Ray Jackendoff, Alex Lascarides, Mark Liberman, Per Linell, Andy Lücking, Aliyah Morgenstern, Catherine Pelachaud, Laurent Prévot, Geoff Pullum, Hannes Rieser, Mark Steedman, Ye Tian, and Alessandro Zucchi for their comments on an earlier draft and for many useful suggestions.

We acknowledge support by the French Investissements d'Avenir--Labex EFL program (ANR-10-LABX-0083) and by the Disfluences, Exclamations, and Laughter in Dialogue (DUEL) project within the Projets Franco-Allemand en sciences humaines et sociales of the Agence Nationale de Recherche (ANR) and the Deutsche ForschungGemeinschaft (DFG) and, for its inspiring atmosphere, the café in the Israel Museum, Jerusalem.
The notion of Interaction Situation is a generalization of the notion of ‘Discourse Situation’ introduced by Barwise & Perry (1983).

‘Hello’ has an ironic use (‘Are you at all plugged in?’) that is not conversation initial, a use that has also become conventionalized, as indicated by the fact that ‘Hi’ lacks this use.
For additional data about such cross-turn presuppositions see (Linell, 2009; Linell & Mertzlufft, 2014) on the Swedish and German x-och/und-x and initial double auxiliary constructions.

The term ‘own communication management’ is due to Jens Allwood, see e.g., (Allwood et al, 2005) for discussion.
See also data from http://itre.cis.upenn.edu/~myl/languagelog/archives/003011.html and from (Yuan, Liberman, & Cieri, 2006), showing the regularity of the extent of speech events.
With respect to (28d), it’s unclear whether repetition of a gesture accompanied by a quizzical face conveys a clarification request; Catherine Pelachaud (p.c.) has suggested to us that it might; this is currently the subject of an experimental study. However, at least in corpora where gesture clarification has been studied, one apparently finds only examples like (28b), as in the corpus described in (Healey et al., 2015), data we thank Nicola Plant (p.c.) for.

Specifically, they seem to refer to Broca’s area—see (Jiang et al., 2012), Figure 1.
That is, inference whose conclusion is, respectively, a proposition or a question—we exemplify both kinds of inference below.
Extensive motivation for this can be found in (Fernández, 2006; Ginzburg, 2012), based primarily on semantic and syntactic parallelism in non-sentential utterances such as short answers, sluicing, and various other fragments.

Thus, the FEC in the QUD associated with a wh-query will be the wh-phrase utterance, the FEC in the QUD emerging from a quantificational utterance will be the QNP utterance, whereas the FEC in a QUD accommodated in a clarification context will be the sub-utterance under clarification.
In the limit, no roles are queried and the question is a polar one, posed to confirm the intended content:

• A: Do you like Hrvati? B: Do I like Hrvati?

• MaxPending utterance content for A: Ask(A,?like(B,h))

• Content of B’s clarification question: ?Ask(A,?like(B,h)) ("Are you asking if I like Hrvati")

As Ray Jackendoff (p.c.) reminds us, this need not hold for negative (‘She never said ‘ . . . ’”) or hypothetical (‘If I ask ‘ . . . ’”) direct quotations. This is an instance of a more general interaction between negation and conditionalisation and event reference.
Also relevant in this respect are pivot constructions discussed in (Norén & Linell, 2013); frequent in conversation, neither self–, nor other–corrected, violating basic selectional principles: (i) E: oh that’s what I’d like to have is a fresh one. ((Norén & Linell, 2013), example (1)).
The meanings of events of uttering *the latter* N, *the first* N, . . . *the last* N, *the former one*, etc. are specified in a similar way.

7 References


Steedman, M. (2013). Romantics and revolutionaries: What Theoretical and Computational Linguists Need to Know about Each Other (and were ashamed to ask). Linguistic Issues in Language Technology, 6(11).


