MULTIMODAL COMMUNICATION IN LFG: GESTURES AND THE CORRESPONDENCE ARCHITECTURE

Gianluca Giorgolo and Ash Asudeh
Carleton University
Carleton University & University of Oxford

Proceedings of the LFG11 Conference

Miriam Butt and Tracy Holloway King (Editors)

2011

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

In this paper we investigate the interaction between verbal language and the non-verbal behaviours that commonly accompany it. We focus on spontaneous hand gestures. We discuss the complex network of interactions between the two modalities and how we can model the interpretation of a multichannel signal. We embed this model of interpretation in the LFG correspondence architecture and we show how the flow of linguistic information that characterizes the architecture can be used to make the interpretation more precise. The result is an enriched architecture in which a complex signal is first broken up into its component parts. The subcomponents are initially interpreted independently but are then fused at the end into a single meaning object. Our model can capture quite precisely the intuitive meaning associated with multimodal utterances.

1 Introduction

In this paper we take a step back from the intricacies of the grammar of natural language and look at it together with the non-verbal behaviours that, more often than not, accompany it. In particular, we examine the spontaneous manual gestures that are produced universally in connection with verbalizations. The goal of the paper is to show how this behaviour is actually very much connected to the complex grammatical structures of natural language and how we can capture these relationships in the framework of the correspondence Architecture of Lexical-Functional Grammar (LFG; Kaplan and Bresnan 1982, Bresnan 2001, Dalrymple 2001). Our claim is that the correspondence architecture (Kaplan, 1987, 1989; Asudeh, 2006, 2012) is an ideal model to represent the interactions between the verbal and the gestural modalities, given the possibility of controlling, at a very fine-grained level, the flow of information between different analytical structures.

The fact that spontaneous gestures play a role in *conveying information* together with verbal language is nowadays well supported by a growing body of studies. Gesture is not a primary mode of communication, and yet the information conveyed solely in this modality is quite consistently integrated in the mental models of reality that we create during a face to face conversation. The first studies of gestural behaviour, in particular the seminal work of Adam Kendon and David McNeill (Kendon, 2004; McNeill, 1992), already stressed that veridical information that is not verbalized is present in the mental representations of participants of a conversation. This observation has been confirmed over the years by a number of behavioural (Kita, 2000; Kita and Özyürek, 2003; Giorgolo, 2010) and neuropsychological experiments (Özyürek et al., 2007; Willems and Hagoort, 2007).

[†] This research is supported by an Early Researcher Award from the Ontario Ministry of Research and Innovation and NSERC Discovery Grant #371969. The first author would like to thank Hannes Rieser for allowing access to the Speech and Gesture Alignment Corpus. The authors thank Doug Arnold, Dag Haug, and the audience at LFG11 for their comments and questions.

Another important characteristic of the interaction between language and gesture that emerges from the data collected in the field and the lab is the fact the two modalities are not simply paired in an unrestricted way, but instead that there are *constraints* on how gestures and language can co-occur. The constraints, that apply both to the production (Kita et al., 2007) and the perception (Giorgolo and Verstraten, 2008) ends of communication, cut across the classical levels of analysis of natural language. Therefore gesture and language are, in McNeill's terminology, simultaneously *synchronized* along different dimensions of analysis (McNeill, 1992):

1. At the prosodic level we observe a strict relationship between *pitch* and *amplitude* peaks (and in general stress patterns) and the *stroke* of a gesture, the most effortful and kinetically prominent phase of a gestural action (Loehr, 2007; Giorgolo and Verstraten, 2008).

2. The alignment between prosodic peaks and gestures' strokes has a clear effect on the overall temporal alignment between gesture and speech, in particular with respect to syntactic constituents and their interpretation: gestures are temporally aligned with the linguistic expressions they are informationally related to.

3. Temporal alignment is in a sense also a form of semantic alignment, as the information conveyed by the gesture must be compatible with the interpretation of the linguistic expression they accompany (i.e. gestures cannot negate information that is expressed verbally (Lascarides and Stone, 2009)); there is however another sense in which gestures are semantically aligned with language: there are in fact limitations to the distribution of the semantic "constituents" gestures can accompany. In particular, gestures seem to behave as modifiers of first order properties/relations; we return to this point below.

4. Finally, at the level of discourse and information structure, we see that gestures are sensitive to linguistic patterns; for example they align with anaphoric relations by re-offerring related manual representations accompanying the linguistic expressions that take part in the relation.

The fact that the data about gestures so strongly suggests a fundamental role of simultaneous alignment patterns in determining the "grammaticality" of gestures motivates our choice of using the correspondence architecture to jointly model gesture and verbal language. In fact, at a sufficient level of abstraction, the correspondence architecture is a model of alignment, as the different structures hypothesized by LFG can be interpreted as *simultaneous constraints* that jointly direct the interpretation of a linguistic expression. With a physical metaphor we could interpret the linguistic expression as a complex signal built up by the composition of synchronized more elementary signals (the various structures). Then the interpretation of the expression becomes a process of decomposion of the signal in its subparts that together allow us to estimate its source (the meaning of the expression). Our idea is to extend the process to include the input coming from an additional synchronized modality.

In this paper we will focus on the interaction between language and gesture at the syntactic and semantic levels. We will demonstrate how we can use the correspondence architecture to capture the joint contribution of speech and gesture to interpretation and how we can use the rich grammatical information associated with linguistic syntactic structure to make more precise the massively ambiguous meaning that we can attach to a gesture in isolation. For this demonstration, we will analyze some general properties of gestures and show for a particular example how a grammatical feature like NUMBER can restrict the space of possible meanings of a gesture.

In Section 2 we introduce some background notions on gestures and on the theory of gestural interpretation presented by Giorgolo (2010), which we use as a basis for our analysis. Section 3 discusses the details of the integration of an additional expressive modality to the correspondence architecture and how the interpretation process must be modified to generate a single joint meaning object. Section 4 explores the implications of our proposal by analyzing in depth an example from the Speech and Gesture Alignment (SaGA) corpus (Lücking et al., 2010), an annotated multimodal corpus of diadic interactions. We conclude in Section 5 with some final remarks.

2 Background: Iconic Gesture

For reasons of space, we will concentrate our discussion about multimodality to a class of gestures known in the literature as *iconic gestures*. An example of this type of gestures is shown in Figure 1. The example is extracted from the SaGA corpus (Lücking et al., 2010).[1] The gesture accompanies the utterance *und hat zwei Türme* 'and has two towers', describing a church with two towers. The stroke of the gesture temporally overlaps with the DP *zwei Türme*, and it provides a visual representation of the spatial extension of the two towers referred to by the verbal expression.

2.1 Properties of Iconic Gestures

This example allows us to present some of the key properties of iconic gestures. The first key property of iconic gestures illustrated by the example is the type of information they normally convey. The gesture under discussion provides a visual representation of the physical/spatial properties of the towers, such as their

[1] The SaGA corpus was collected with German speakers and therefore all the examples in the paper will be in German. However all our generalizations are intended to be extended also to other languages. We decided to use naturally occurring data to stress that the study of such a subconscious activity as spontaneous gestures requires the use of empirical data to be study successfully.

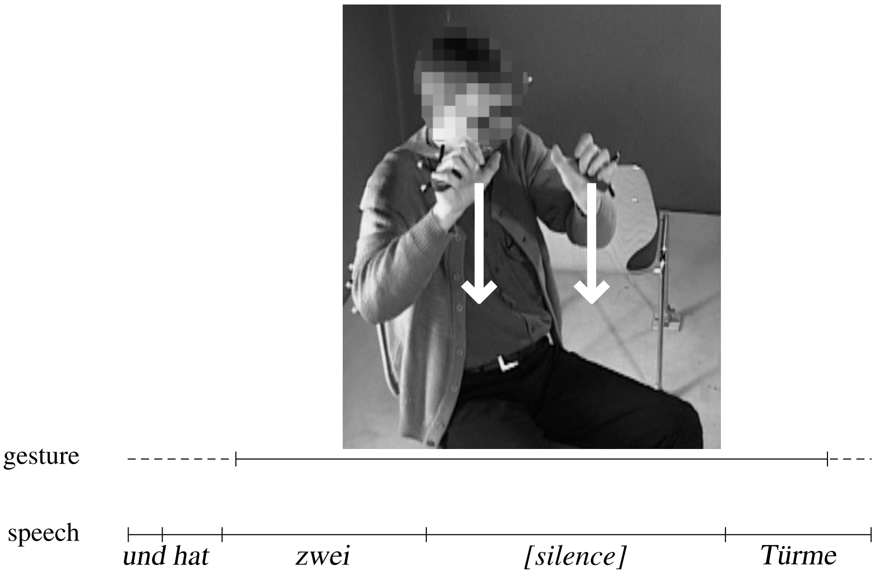


Figure 1: Example iconic gesture.

relative position, their orientation, the fact that they are disconnected, and we are also given a rough approximation of their shape. Iconic gestures generally convey information that is spatio-temporal in nature, as they normally describe properties of physical referents and events. They differ in how this information is conveyed; for example this gesture creates a sort of miniature image of the towers, while a gesture describing an action performed by a human being will normally take the form of an enactment of the action, giving us an internal perspective on it. However the information can always be modeled as specific regions of a spatio-temporal frame of reference.

Another interesting property of iconic gestures illustrated by the example is the way in which they are generated. The gesture shown in Figure 1 is created on the spot by the speaker, possibly on the basis of the mental imagery that the speaker has of the referent he is describing. In general iconic gestures lack a conventionalized form and in this sense they are different from those gestures that have a fixed meaning inside a speech community. Iconic gestures manage to convey meaning solely by the fact that they somehow *resemble* their referent. This fact will be quite relevant in the choices we will make when modeling iconic gestures in LFG, as the lack of a conventionalized form, and consequently of an agreed upon meaning, prevents us from treating them as regular lexical resources. Our solution will be to associate with gestures a very general (i.e. underspecified) lexical entry, constructed only on the basis of the properties that are observable from their formal appearance.

This last choice is also motivated by the fact that the interpretation of iconic gestures is massively dependent on contextual factors, in particular on the linguistic context in which they are embedded. The interpretation of iconic gestures becomes in fact almost impossible without an accompanying verbal expression. The only information obtainable is, as stated above, the bundle of spatial properties associated with the virtual space created by the gesture. This reflects a more general limitation of the possibility of conveying information via the gestural channel. As we will see below the semantic function that a gesture has is restricted to a form of intersective modification of first-order properties. A gesture imposes additional constraints (of a spatial nature) on the set of referents identified by a property. Other functions, such as the introduction of new referents, the independent introduction of a negative polarity context or the creation of a predicate-argument structure, are beyond the semantic expressivity of gestures. Gestures rely on the logical structure set up by verbal language and simply operate inside these logical structures without modifying them. The semantic contribution of gestures is therefore comparable to that of content words.

2.2 Interfacing Gesture and Language

With this information in the background, we now move on to analyze how the two modalities collaborate in conveying a conjoined meaning. To answer this question we first need to address two subquestions. The first one concerns the interpretation of gestures as isolated objects. Iconic gestures never occur outside of a speech fragment; nevertheless their interpretation must first go through an independent interpretation step, given that the processing of the activity of the hands is not in any way connected to the processing of verbal language. The second question concerns the fusion step of the interpretation process: once we have associated with a gesture a (largely ambiguous) interpretation we must specify how this information is combined with speech, keeping in mind the multiple constraints coming from the different levels of alignment.

To give precise answers to these questions we use the formal framework for the analysis of gestures introduced in Giorgolo (2010). The framework consists in an extension of classical Montagovian semantics together with a formal logic designed to describe space and time. With these ingredients we can be very precise about the process of interpretation of a multimodal utterance.

The answer to the first subquestion is based on the representational characteristic of gestures and their communicative function. We take a gesture to convey a type of information that we can model as an equivalence class of spatial objects that are informationally indistinguishable from the virtual space set up by the hands. The equivalence part of the meaning is contributed by the *representational semantic function* of the gesture: a representation in general does not refer necessarily to a single instance but rather it can refer to all objects and events that are similar (in a way to be made more precise) to the physical appearance of the representation. The specific equivalence class and the level of informativity is instead provided by the actual formal properties of the gesture. Giorgolo (2010) introduces a family of description logics that are used to match the expressive power observed in iconic gestures. Each logic is not a single language, but rather a family of related languages. This is motivated by the following considerations:

Modularity. Certain spatial properties are necessarily preserved by iconic gestures. Other spatial properties may be disregarded. For instance a gesture may give us a faithful representation of the relative position of different entities, such as when we draw a virtual map for our interlocutor, but the precise shape of these objects is usually largely left unspecified (they could be for instance just amorphous blobs). We need a modular language in which we can selectively add or remove predicates that are associated with specific spatial properties (e.g. orientation predicates, position predicates, shape predicates, etc.). Most importantly, these predicates should be independent of each other as we need to be free to fine tune the logic according to what we observe in the gesture (however, see Giorgolo (2010) for a discussion of a number of possible interdependencies among different groups of properties).

Simplification. Consecutive gestures that refer to the same entity or event follow a pattern of decreasing informativity. The sets of spatial properties that the subsequent gestures conserve are ordered by a subset relation. So, for instance, the gesture shown in Figure 1 is repeated by the speaker two other times later in the conversation, when referring back to the same church. In both cases we observe a decrease in the amount of visual information expressed in the gesture. In the first repetition the speaker drops the depiction of the three dimensional shape of the towers, while the fact that they are disconnected and that they are vertical is still depicted. In the last repetition, the only information available seems to be that the towers are two in number, as the gesture resembles the conventionalized gesture for the number two. This pattern mirrors quite closely the tendency in language to consecutively refer to entities and events in more economic/simpler ways (e.g., *The man who Thora saw yesterday ... the man ... he*).

Specifically, we use a family of languages based on a theory of region-based space-times to reproduce the third-person perspective we observe in the gesture of Figure 1, and another family of languages based on a theory of human gestural articulators (e.g., fingers, hands, arms, joints) to represent the embodied perspective typical of gestures representing actions. In this way we can represent the informational content of a gesture as the collection of the proposition in the chosen description logic that are satisfied in the virtual space set up by the gesture, what we will call the *theory* of the gesture. The interpretation of a gesture in isolation will then correspond to the characteristic function of the equivalence class of spaces that are models for the theory of the gesture. For instance, in the case of the gesture in Figure 1 we first select an appropriate description logic (in this case the third-person perspective one) and create a theory by checking all the spatial properties involving the two regions depicted in the gesture. The theory is the collection of all propositions (positive for the present properties, and negative for the absent ones) that are satisfied in the space under consideration. In our case, the collection would include a proposition stating that the two regions are disconnected, that they are vertical, that they are *not* one above of the other and so on. The interpretation we assign to the gesture corresponds to the set of all spaces made up of two regions that also possess the spatial properties (both positive and negative) encoded in the theory.

We now move on to the second question, the one about the integration of the two modalities. As already stated, gestures cannot introduce novel referents, nor can they change the polarity of the context in which they appear, the only function they can perform is to place additional constraints on the interpretation of the referents and the events already introduced by language. This suggests a semantic function akin to the one of *intersecting modifiers*. Therefore we propose to reduce the interface between the two modalities to a generalized form of *intersection*. To obtain this generality we assume that the semantic toolkit at out disposal includes a collection of boolean algebras for all the boolean types. This is actually a rather inexpensive assumption, as the same process is necessary in language to model the cross categorial behaviour of conjunctions. We can therefore consider this logical operation to be one of those available in general in communication. Intersection is implemented as the *meet* operation of each boolean algebra. This allows us to have a flexible notion of intersection, because the same gesture can combine with constituents of different semantic types, as shown indirectly by Alahverdzhieva and Lascarides (2010). At the same time we predict that gestures combine only with semantic constituents with the appropriate type. In fact, beside excluding any non-boolean expression from the set of possible linguistic correlates of a gesture, the meet operation also requires the two semantic expressions to be of the same type. We will see in the next paragraph that we relax this requirement to a form of equality under a homomorphic mapping, but the meaning terms that are intersected are required to have the same "arity". This requirement is sometimes too strict, as there are cases in the data in which we want to combine objects that *prima facie* have different arities. In all these cases it seems that linguistic factors influence the integration of the modalities by providing clues for the adaptation of the gestural interpretation. The Correspondence Architecture allows us to model these effects elegantly and in Section 4 we will see how a grammatical feature can be used to resolve such a type-clash situation.

2.3 Multimodal Interpretation

At this point we are ready to describe in detail the process of interpretation for a multimodal utterance. We give a graphical representation of the process in Figure 2. The diagram describes the process by which a single gesture and a verbal language fragment are first independently interpreted and how their interpretations are then joined into a single one. Γ and Σ respectively represent the gesture and the language fragment. The verbal expression, Σ, is interpreted by a standard interpretation function, $[\![\cdot]\!]_f$, yielding values taken from a frame of reference, F. F is a collection of domains of the usual kind, built on top of an ontology of entities e, truth-values t and events s. The frame, F, is related to a spatial frame of reference, S, by a family of (possibly partial) functions, Loc, which mirrors the compositional structure of F into S. S is a set of domains constructed in a way similar to F: we start from a set of primitive types and then we inductively define the remaining types as those corresponding to all the functions whose domains and codomains are the primitive and the derived types. In the case of S the primitive types will be *regions* r, truth-values t and events s. The types of F (i.e. e, t, s) are then mapped through members of Loc to the types of S according to the following conditions (where loc_a is the specific member of Loc mapping objects of type a to objects in S):

1. $loc_e(x) = y$, where y is of type r
2. $loc_t(x) = x$
3. $loc_s(x) = x$
4. $loc_{a \to b}(f) = g$, such that for all objects x of type a we have that:

$$g(loc_a(x)) = loc_b(f(x)) \ .$$

In other words, Loc identifies a homomorphic image of the traditional abstract interpretation of the speech signal in the spatial domain and specifies how the spatial interpretation is constructed from the abstract frame of reference obtained from the speech signal.

The composition of the interpretation function from Σ to F and Loc therefore defines a interpretation function, $[\![\cdot]\!]_s$, from Σ directly to S. The composition may not always be defined, as we do not require every verbal expression to have a spatial extension (e.g. logical words like determiners, modals and conjunctions lack a direct spatial interpretation, although they may have a metaphorical one). The distribution restriction of iconic gestures allows us to be sure that the interpretation process will never require us to access the spatial extension of those expressions. On the left side of the diagram, ω maps from a collection of features representing the gesture to a representational space, RS. ω takes into account various constraints, such as the mode of representation (drawing, sculpting, shaping, enacting, etc.) and deformations of the gestural space due to physiological constraints. Finally, the representational space, RS, corresponding to the gesture and the spatial representation, S, of the speech signal are combined by requiring an informational equivalence, such that they must satisfy the same set of spatial constraints. The combination is implemented as the meet operation. The meaning of the verbal expression becomes intersectable with the meaning of the gesture thanks to its transformation via the Loc mappings.

In the next section we show how we propose to embed the interpretation process just described in the correspondence architecture. To do so we will need to

$$\begin{tikzcd}
\text{Gesture Structure }(\Gamma) \arrow[dr, "\omega"] & & \mathrm{F} \arrow[d, "Loc"] & \text{Phonological String }(\Sigma) \arrow[l, "{[\![\cdot]\!]_f}"'] \arrow[dl, dashed, "{[\![\cdot]\!]_s}"] \\
& \mathrm{RS} & \mathrm{S} \arrow[l, leftrightarrow, "\equiv"] &
\end{tikzcd}$$

Figure 2: Interpretation process for a multimodal utterance.

accommodate the gestural component in the architecture. But the model will also be enriched by the rest of the architecture. We will see how access to the other structures created during the interpretation process in the LFG architecture can be used to improve on the predictions made by the model in its current form.

3 Integration of Gesture in the Correspondence Architecture

In order to extend the LFG framework to deal with multimodal utterances, we introduce certain modification to the correspondence architecture. The new version of the architecture is shown in Figure 3. This version of the correspondence architecture is based on the pipeline version of the standard architecture, which is discussed by Bögel et al. (2009) and Asudeh (2012).

The first modification is to assume that the **Form** end of the pipeline is a multimodal utterance, rather than a phonological string. The linguistic part of this utterance is then mapped to the phonological string by the υ correspondence function.

Parallel to the υ function we introduce the γ correspondence function. The γ function maps the multimodal utterance to a timed stream of *gesture structures*. Each gesture structure is simply a feature structure describing the physical appearance of the gesture (typical features include hand shape, trajectory, orientation, and so on).

The third modification is to define a level of time structure, whose purpose is to align gestural elements and linguistic elements. Time structure is a time-indexed set of the substrings in the phonological string. The time structure is populated by a function τ from the phonological string. These time indexes are then propagated to the constituent-structure, resulting in a tree whose nodes are time indexed. The correspondence function κ specifies in the time structure the substrings that are temporally aligned with different gesture structure in the gesture stream by creating links between gesture structures and those substrings with which the gesture is synchronized. We assume that two types of links are established, depending on the nature of the linguistic context in which the gesture appears. The first type

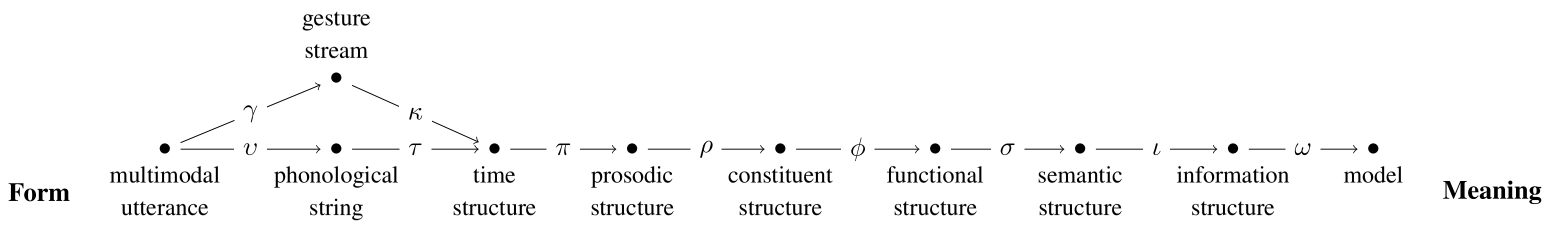
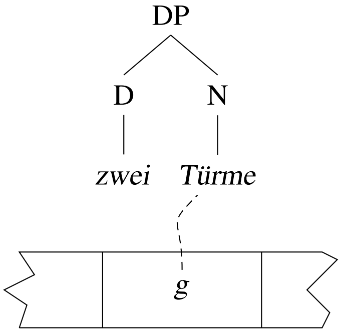


Figure 3: Correspondence architecture, modified pipeline version



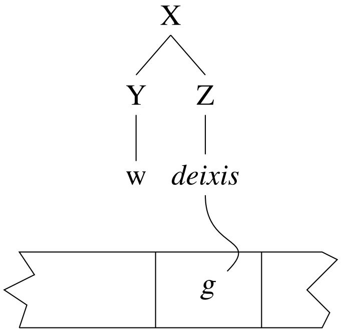
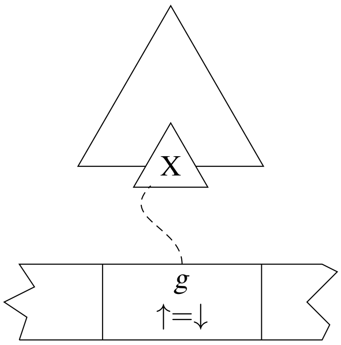


Figure 4: Temporal links between gesture stream and constituent structure.

of link correspond to the case of gestures that are simply performed in parallel with language, while the second one corresponds to the case of those gestures that are "marked" explicitly in language via some form of deixis (e.g. *I caught a fish* <u>*this*</u> *big*). Figure 4 shows how these links can be propagated in the architecture to create links between the gesture stream and the constituent structure. The dashed line represents the first type of links while the continuous line the second type. In the rest of the paper we will concentrate only on the first type of link.[2]

The last modification we propose is in the way the functional-structure is generated. In the language-only case, the functional structure is the result of applying the ϕ mapping to the constituent structure. In our case we need to make the ϕ map aware also of the gesture structures contained in the gesture stream and of the links defined between the time-structure and the gesture stream. All this information is of course available to the ϕ structure (given the pipeline shape of the architecture) and simply requires the introduction of a rule that determines how a gesture contributes to the functional structure of the multimodal utterance. We give here a graphical representation of the rule:



The rule consists in a functional constraint saying that the functional structure of a gesture g (see below for typical functional structures of gestures) is the same

[2] The case of links generated for deictic elements is actually trivial once we have defined the first type of link, but requires in depth discussion of the linguistic elements that trigger this type of link. We leave this discussion for future work.

as the one of the node X that it is linked to, obtaining the same effect of the familiar constraint $\uparrow=\downarrow$. To maintain a uniform notation in our functional constraints, we will use the abbreviations $\downarrow$ and $\uparrow$ also for the multimodal links: $\downarrow$ will refer to the functional structure of the gesture, while $\uparrow$ will be used to refer to the functional structure associated with the node to which the gesture is linked.

Finally, the ω correspondence function completes the mapping from the bundle of kinetic, physical features to the representational space. Since ω is late in the Form-Meaning pipeline in the modified correspondence architecture, it can also be sensitive to information earlier in the pipeline, particularly functional structure information. Information extracted from the functional structure can be used to appropriately instantiate the meaning of the gesture such that it takes into account morphosyntactic properties of its linguistic correlate.

In the next section we provide an in depth analysis of how a multimodal utterance is interpreted in our revised architecture.

4 Analysis

To demonstrate the advantages offered by the projection architecture in modeling the integrated interpretation of gesture and speech, we reanalyze an example presented in Giorgolo (2010), which is extracted from the SaGA corpus. The example is the one presented in Section 2. The speaker is describing a church with two towers and accompanies the utterance of the DP *zwei Türme* ‘two towers’ with a gesture depicting some spatial information about the towers. The gesture gives us information about the relative position of the towers (they are parallel) and about the fact that the towers are disconnected. We are also given a rough representation of the shape of the two towers, two vertically oriented prisms. We now follow the interpretation process depicted in Figure 2 and see how the various components of our revised correspondence architecture contribute to produce the final meaning of the expression.

The multimodal utterance is split by the υ and γ maps into its component parts. The gesture stream in this case is composed of a single gesture structure. The gesture structure is generated by the γ function from the raw, visual data (in our case the role of the γ function has already been played by the team of annotators that created the corpus). A partial representation of the resulting functional structure is shown in Figure 5.

The phonological string is mapped to a time structure and a link is created between the gesture and the substring it is related to. In our case we have two choices, depending also on the status we attribute to the word *zwei* ‘two’. If we consider the numeral a determiner (possibly the most conservative of the two options), then, given the distributional restriction on gestures we are forced to link the gesture structure to the substring *Türme*, as the quantified phrase *zwei Türme* is of too high an order for a gesture (being a property of properties). The other option is to consider the numeral a form of intersecting modifier: in this case we are free to link

$$\begin{bmatrix} \textsc{Left.HandShapeShape} & \text{loose C} \\ \textsc{Left.PathofHandShape} & 0 \\ \textsc{Left.HSMovementDirection} & 0 \\ \textsc{Left.HandShapeMovementRepetition} & 0 \\ \vdots & \vdots \\ \textsc{Right.HandShapeShape} & \text{loose C} \\ \textsc{Right.PathofHandShape} & 0 \\ \textsc{Right.HSMovementDirection} & 0 \\ \textsc{Right.HandShapeMovementRepetition} & 0 \\ \vdots & \vdots \end{bmatrix}$$

Figure 5: Partial gesture structure.

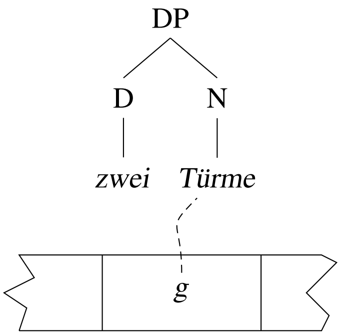


Figure 6: Links between the gesture stream and the DP *zwei Türme*.

the gesture either to substring *Türme* or to *zwei Türme*. Giorgolo (2010, p. 65) shows that in similar cases the resulting interpretations are not truth-functionally distinguishable. Both choices are motivated by the temporal alignment we observe, as strokes are not perfectly aligned with their linguistic correlates (they can "leak" over other elements and have some freedom of movement inside a specific time-window). We choose the first link point, in order to avoid having to include an existential closure operation to bind the plural tower referent. The result is the linking structure shown in Figure 4. However notice that in this case, had we made the other choice, the final interpretation would have been the same. Our model is therefore not capable of distinguishing the two choices at the truth functional level. This could be a limitation of our proposal but it could also reproduce a real indeterminacy and a limitation of the contexts with which a gesture can compose. The answer to this question requires an in depth analysis of the distribution of gestures according to compositional parameters, a task we leave for future work.

The gesture and the noun it is linked to are defined by the linking rule to map to the same functional structure. The gesture generally does not add functional

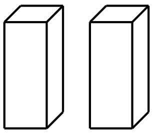


Figure 7: Virtual space generated from the gesture structure.

structure information, but uses information in its functional structure to constrain interpretation and potentially places constraints on the functional structure it contributes to. The resulting structure for the DP is shown in (1). As we can see, it is same one we would obtain without considering the gesture.

(1)
$$\begin{bmatrix} \text{PRED} & \text{`tower'} \\ \text{NUMBER} & \text{PL} \\ \text{SPEC} & \begin{bmatrix} \text{PRED} & \text{`two'} \end{bmatrix} \end{bmatrix}$$

As stated above, iconic gestures lack a conventionalized meaning: they are not lexicalized. However we can associate with them a lexical entry that is directly obtainable from the formal features of the gestures, as they are described in the gesture structure. To generate the lexical entry for a gesture we need to extract some information from the gesture structure. We interpret the description of the gesture in the gestural structure as the input for a constraint resolution problem that in the end generates a spatial configuration that corresponds to the virtual space set up by the hands. Specifically, features like hand shape, direction of movement, etc., allow us to determine the number, location and shape of the regions involved in the gestural representation. This information can then be used to generate the functional constraints and the semantic terms that make up the lexical entry of the gesture. In the specific case we are considering here, we can see that from the feature structure we generate a space like the one represented in Figure 7. The space generated in this way does not correspond yet to the core meaning of the gesture. We need to extract from it the spatial information that allows us to define the equivalence class forming the meaning of the gesture in isolation. To do that, we generate the theory of the gesture, as described above, by taking the set of propositions of the desired spatial logic that are true in the virtual space.

At this point we are ready to discuss the lexical entry for a gesture of the type of our example. A partial general lexical entry is shown in (2). The representation is partial in the sense that the disjunction should be extended to deal with additional structures whose interpretation corresponds to a binary relation or to a property of entities composed by two sub-elements. Alternatively we could introduce variables for propositions in the glue logic terms.

(2) g $(\uparrow \text{NUMBER}) \neq \text{PL}$

$$\lambda R.\lambda x.\lambda y.R(x,y) \wedge core(loc_e(x))(loc_e(y))$$

$$((\uparrow \text{ OBJ})_\sigma \multimap (\uparrow \text{ SUBJ})_\sigma \multimap \uparrow_\sigma) \multimap$$

$$((\uparrow \text{ OBJ})_\sigma \multimap (\uparrow \text{ SUBJ})_\sigma \multimap \uparrow_\sigma)$$

$$\vee$$

$(\uparrow \text{NUMBER}) =_c \text{PL}$

$$\lambda P.\lambda x.P(x) \wedge (\delta(core))(loc_e(x))$$

$$((\uparrow_\sigma \text{ VAR}) \multimap (\uparrow_\sigma \text{ RESTR})) \multimap$$

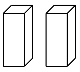
$$((\uparrow_\sigma \text{ VAR}) \multimap (\uparrow_\sigma \text{ RESTR}))$$

The general shape of the entry is suitable for any iconic gesture depicting two distinct regions. The entry lacks a syntactic category, as gestures do not take part in any grammatical function but are merely a reification of meaning. Notice that even in those cases in which language marks through deixis the necessity of interpreting the gesture to obtain a full interpretation, gestures are not necessary to determine the grammaticality of the verbal utterance.

The semantic part is composed by a disjunction of possible interpretations. This models the strong ambiguity of a gesture outside of a linguistic context. To reflect the necessity of linguistic information to disambiguate the meaning of a gesture we use functional constraints and the shape of the glue logic terms (Dalrymple, 1999; Asudeh, 2012) to select for a specific interpretation. In the case under consideration the two interpretations presented here can be distinguished by the feature NUMBER of g's functional structure. The idea is to distinguish between two possible interpretations of the two regions depicted by the gesture. The two regions can in fact be considered as two independent entities related in some way made precise by language, or they could be the discontinuous spatial extension of a single entity, either a plural entity composed of continuous sub-entities or a singular inherently discontinuous entity. As the referent for the gesture is introduced in the linguistic expression, in our case the variable bound by the determiner *zwei*, we use the grammatical information at our disposal to distinguish between the competing interpretations. The first interpretation presented is selected on the basis of a negative constraint on the feature NUMBER. This interpretation should be selected in case the gesture accompanies a transitive verb. In this case we require the related object not to have a plural NUMBER feature. A verb's f-structure satisfies this constraint, because it is only arguments to verbs, not verbs themselves, that are specified for NUMBER. In fact, we could obtain the same result with a constraint of the type $\neg(\uparrow \text{NUMBER})$. In the case of the second interpretation we use a constraining equation to ensure that the gesture combines with a set of entities whose elements are plural objects. A third interpretation, which we do not discuss here, would require an argument of the linked verbal element to have a singular NUMBER feature, and would give rise to the interpretation that combines the two regions into a singular discontinuous entity.

The two glue terms reflect these distinctions. In the first case we assign to the gesture a semantic function similar to the one of a verbal modifier. The gesture consumes a resource corresponding to a transitive verb and returns the same type of resource. In the second case the gesture acts as a nominal modifier, consuming a first order predicate and returning a new predicate of the same type.

The lambda terms give us the details of how the information contributed by the gesture obtains the modification effect. The two terms of course reflect the different nature of the elements on which they operate. However their general shape is comparable and the gesture-only contribution is identical in the two terms. The core meaning of the gesture is represented by the function $core$, which is a shorthand for the function presented in equation (3).

$$core = \lambda r_1.r_2.\,(r_1 \cup r_2) \equiv \tag{3}$$


The core meaning of the gesture is a boolean function, taking two regions (of type r) as arguments and returning a truth-value. The two regions are combined in a single space via a sort of union operation and the resulting space is then required to be a model for the theory of the gesture that we represent synthetically as the figure in the righthand side of the equivalence. In other words, the function checks if the space composed by the two regions passed as arguments is similar to the one represented by the gesture. This function corresponds to the equivalence class of spaces of which the gesture can be a representation. In this case the equivalence class defined by the theory of the gesture corresponds to the set of spaces composed by two distinguished regions that are disconnected, that are parallel, whose main axis is vertical and whose shape is of two prisms.

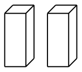
In the case of the first interpretation, the arguments to the $core$ function are simply the spatial projections (i.e. the image under loc_e) of the two referents corresponding to the object and the subject of the transitive verb. The boolean result of the function is then "met" with the application of the binary transitive predicate to the same referents.

In the second case, the two arguments are obtained by using a *distributivity* operator δ, defined in equation (4), that splits a plural entity into its atomic parts (in our case the plural towers are decomposed into the singular towers) and then passed to the $core$ function. Also in this case the result of the application is met with the meaning provided by verbal language.

$$\delta(x) = \lambda e.x(e_1\ \cdots e_n) \tag{4}$$

Given the functional structure associated with *Türme* we select the second interpretation. The resource offered by the gesture enters the glue proof in the same way as standard lexical items (i.e. as an axiom) and the resulting proof term is the one shown in (5). The term describes a function from first order properties to truth values. The argument Q represents the scope of the quantified phrase *zwei Türme*. The determiner *zwei* introduces the existential quantifier and the condition on the variable x to be assigned a plural entity with cardinality 2. The predicate $tower$ is contributed in the usual way by the noun *Türme*. The rest of term is contributed by the gesture and corresponds to the condition imposed on the existentially quantified variable x by the manual representation. Specifically, the spatial extension of the referent should be a plural object decomposable into its composing regions (which should be two) and such that the two regions are disconnected, they are parallel, their main axis is vertical and their shape is roughly that of a prism.

$$\lambda Q.\, \exists x.\, Q(x) \;\wedge\; |x| = 2 \;\wedge\; tower(x) \;\wedge \tag{5}$$

$$(\delta(\lambda r_1.r_2.\, (r_1 \cup r_2) \equiv$$  $$))(loc_e(x))$$

This interpretation corresponds to the intuitive meaning that we would associate with the gesture under consideration in this linguistic context.

5 Conclusion

In this paper we have investigated the nature of the relationship between verbal language and the non-verbal behaviours that commonly accompany it. We have focused on spontaneous iconic gestures and discussed how the interaction between the two modalities is not restricted to a simple pairing of different communicative channels, but rather follows a number of complex rules. The interaction is based on constraints on the temporal and prosodic alignment between the two modalities but also on deeper connections that include interactions between gesture and language at the morphosyntactic and semantic levels.

The goal of the paper was to approach multimodal communication from the perspective of LFG's correspondence architecture. We have demonstrated that we need a rich and fine-grained framework, such as the one offered by LFG, in order to capture the complexities of multimodal communication. We have first presented a model for the interpretation of multimodal utterances based on standard semantic tools and a logical language that matches the representation power observed in iconic gestures. We have discussed how the interpretation is nevertheless dependent on linguistic factors that need somehow to control the creation of meaning. The correspondence architecture offers precisely this possibility thanks to the flow of information between different levels of analysis that allows for an interaction between them.

To integrate multimodal signals in the LFG framework we introduced a number of additions to the architecture, leaving the language-only components basically untouched. One of the main innovations is the introduction of a structure parallel to the phonological string that we called the gesture stream and that represents the temporal sequence of gestures as observed in the multimodal signal. The gestures are represented as feature structures describing their physical appearance. The other fundamental innovation is the introduction of links between the elements of the gesture stream and the nodes of the constituent structure. In this way we are able to let the gesture have access to the functional structure of its linguistic correlate. The information available in the functional structure is used to specify the otherwise largely ambiguous interpretation that we associate with gestures. In particular we have demonstrated how a grammatical feature such as NUMBER can guide the interpretation of a gesture in the desired direction. We envisage that other grammatical features play a similar role in other contexts. For instance a feature like ASPECT can guide the interpretation of the properties of gestures such as repeated similar movements in the case of contexts made up by a verbal phrase. In these cases, ASPECT could allow us to interpret the presence of a repetition as a visual marking that we associate with an imperfective verbal form (e.g. habituality) and therefore constrains the interpretation of the gesture as the depiction of multiple but identical events.

References

Alahverdzhieva, Katya and Lascarides, Alex. 2010. Analysing Speech and Co-speech Gesture in Constraint-based Grammars. In Stefan Müller (ed.), *Proceedings of the HPSG10 Conference*, pages 5–25, Stanford, CA: CSLI Publications.

Asudeh, Ash. 2006. Direct Compositionality and the Architecture of LFG. In Miriam Butt, Mary Dalrymple and Tracy Holloway King (eds.), *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*, pages 363–387, Stanford, CA: CSLI Publications.

Asudeh, Ash. 2012. *The Logic of Pronominal Resumption*. Oxford: Oxford University Press.

Bögel, Tina, Butt, Miriam, Kaplan, Ronald M., King, Tracy Holloway and Maxwell, III, John T. 2009. Prosodic Phonology in LFG: A New Proposal. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG09 Conference*, pages 146–166, Stanford, CA: CSLI Publications.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Dalrymple, Mary (ed.). 1999. *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*. Cambridge, MA: MIT Press.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. San Diego, CA: Academic Press.

Dalrymple, Mary, Kaplan, Ronald M., Maxwell III, John T. and Zaenen, Annie (eds.). 1995. *Formal Issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.

Giorgolo, Gianluca. 2010. *Space and Time in our Hands*, volume 262 of *LOT Publications*. Utrecht: LOT.

Giorgolo, Gianluca and Verstraten, Frans A. J. 2008. Perception of 'Speech-and-Gesture' Integration. In R. Goecke, P. Lucey and S. Lucey (eds.), *Proceedings of the International Conference on Auditory-Speech Perception 2008*, pages 31–36.

Kaplan, Ronald M. 1987. Three Seductions of Computational Psycholinguistics. In Peter Whitelock, Mary McGee Wood, Harold L. Somers, Rod Johnson and Paul Bennett (eds.), *Linguistic Theory and Computer Applications*, pages 149–181, London: Academic Press, reprinted in Dalrymple et al. (1995, 339–367).

Kaplan, Ronald M. 1989. The Formal Architecture of Lexical-Functional Grammar. In Chu-Ren Huang and Keh-Jiann Chen (eds.), *Proceedings of ROCLING II*, pages 3–18, reprinted in Dalrymple et al. (1995, 7–27).

Kaplan, Ronald M. and Bresnan, Joan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pages 173–281, Cambridge, MA: MIT Press, reprinted in Dalrymple et al. (1995, 29–135).

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kita, Sotaro. 2000. How Representational Gestures Help Speaking. In David McNeill (ed.), *Language and Gesture*, Chapter 8, pages 162–185, Cambridge University Press.

Kita, Sotaro and Özyürek, Aslı. 2003. What Does Cross-Linguistic Variation in Semantic Coordination of Speech and Gesture Reveal? Evidence for an Interface Representation of Spatial Thinking and Speaking. *Journal of Memory and Language* 48(1), 16–32.

Kita, Sotaro, Özyürek, Aslı, Allen, Shanley, Brown, Amanda, Furman, Reuhan and Ishizuka, Tomoko. 2007. Relations Between Syntactic Encoding and Co-Speech Gestures: Implications for a Model of Speech and Gesture Production. *Language and Cognitive Processes* 22(8), 1212–1236.

Lascarides, Alex and Stone, Matthew. 2009. A Formal Semantic Analysis of Gesture. *Journal of Semantics* 26(4), 393–449.

Loehr, Daniel P. 2007. Aspects of Rhythm in Gesture and Speech. *Gesture* 7(2), 179–214.

Lücking, Andy, Bergmann, Kirsten, Hahn, Florian, Kopp, Stefan and Rieser, Hannes. 2010. The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In M. Kipp et al. (ed.), *LREC 2010 Workshop: Multimodal Corpora - Advances in Capturing, Coding and Analyzing Multimodality*, pages 92–98.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

Özyürek, Aslı, Willems, Roel M., Kita, Sotaro and Hagoort, Peter. 2007. On-line Integration of Semantic Information from Speech and Gesture: Insights from Event-related Brain Potentials. *Journal of Cognitive Neuroscience* 19(4).

Willems, Roel M. and Hagoort, Peter. 2007. Neural Evidence for the Interplay Language, Gesture and Action: A Review. *Brain and Language* 101(3), 278–289.