Multimodal Communication in LFG:
Gestures and the Correspondence Architecture[*]

Gianluca Giorgolo
Gianluca_Giorgolo@carleton.ca
Carleton University

Ash Asudeh
ash.asudeh@ling-phil.ox.ac.uk
Carleton University &
University of Oxford

July 17, 2011
LFG 2011 · University of Hong Kong

1 Introduction

- Verbal language and gestures co-operate in conveying information in communication. The literature offers evidence from a variety of sources:
  - interaction analysis (McNeill 1992, Kendon 2004)
  - behavioural experiments (Giorgolo 2010)
  - neuropsychological experiments (Willems and Hagoort 2007)
- The interaction between language and gesture is constrained by a number of factors:
  1. Prosody
  2. Temporal alignment, with respect to syntactic constituents and their interpretation
  3. Semantic alignment between information conveyed by gesture and its linguistic correlate
  4. The effects of conversational goals on the distribution of information between modalities
- Alignment is an important notion in multimodal communication. At a sufficient level of abstraction, the Correspondence Architecture is a model of alignment of different sources of information. Therefore, the Correspondence Architecture should also be useful in capturing language-gesture alignment.
- Here we will look a specific kind of gesture, *iconic gesture*, as exemplified in video 1.

[*] This research is supported by an Early Researcher Award from the Ontario Ministry of Research and Innovation and NSERC Discovery Grant #371969.

2 Main Claims

- Iconic spontaneous co-speech gestures interact with language in interesting ways. We here focus on the interaction at the levels of syntactic and semantic structure.
- The Correspondence Architecture allows us to capture a complex network of interactions that involves morphosyntactic features, semantic properties and the formal appearance of iconic gestures.
- Grammatical features like NUMBER can influence the interpretation of gesture.

3 Overview

1. *Introduction*
2. *Main Claims*
3. *Overview*
4. Background: Iconic Gesture
5. Integration of Gesture in the Correspondence Architecture
6. Analysis
7. Conclusion

4 Background: Iconic Gesture

- Iconic gestures have the following key properties:

  1. Iconic gestures are created spontaneously: they lack a conventionalized form. In other words, iconic gestures are not lexicalized.
  2. The interpretation of iconic gestures is massively dependent on the linguistic context; i.e., they are not like pantomime gestures.
  3. Iconic gestures convey information that complements linguistic information by specifying physical/spatial properties of entities and events.
  4. Iconic gestures do not introduce novel discourse referents and do not create a proper predicate-argument structure — a gesture cannot take another gesture or a linguistic element as an argument.
  5. Listeners consistently integrate iconic gestural information.

- To analyze semantic alignment, we need to address two sub-issues:

  1. How is iconic gesture interpreted?
     The interpretation is a collection of spatial properties that can be extracted from the virtual space defined by the hands. Indirectly, this collection of spatial properties defines an equivalence class of spaces that are informationally indistinguishable.
  2. How is this interpretation integrated with the interpretation of the verbal correlate?
     Integration is achieved through *intersection*. That is, integration of gestural information and verbal information is achieved through a similar mechanism to intersective modification in language. This means that gestural information monotonically reduces the space of possible verbal interpretation.

- We use the formal framework of Giorgolo (2010) to make these answers more precise.

  1. How is iconic gesture interpreted?
     Description logics are used to model the representational capabilities of 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. Therefore, we need a modular language in which we can selectively add or remove predicates that express spatial 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. This mirrors 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 spaces to represent a third-person perspective on space and another family of languages based on a theory of human gestural articulators (e.g., fingers, hands, arms, joints) to represent an embodied perspective on space.
  2. How is this interpretation integrated with the interpretation of the verbal correlate?
     In addition to standard semantic tools, such as functions and sets, we assume a 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).

- Figure 1 represents the process of interpretation of a multimodal utterance.
  - Γ and Σ respectively model the gesture and the speech signal.
  - The speech signal, Σ, is interpreted by a standard interpretation function, $\llbracket\cdot\rrbracket_f$, yielding values taken from a frame of reference, F.
  - The frame, F, is related to a spatial frame of reference, S, by a family of functions, Loc, which mirrors the compositional structure of F into S. In other words, Loc identifies a homomorphic image of the 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, $\llbracket\cdot\rrbracket_s$, from Σ directly to S.
  - 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.

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

Figure 1: Interpretation process for a multimodal utterance

5 Integration of Gesture in the Correspondence Architecture

- In order to integrate the gestural contribution with LFG's theory of verbal utterances, we switch to a multimodal utterance perspective, which necessitates certain modifications to LFG's Correspondence Architecture.

- The new architecture is shown in Figure 2 and 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.

- The second modification is to define a level of gesture structure, which the multimodal utterance maps to by the γ correspondence function. A gestural structure is 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. The correspondence function κ specifies in the time structure the substrings that are temporally aligned with elements of gesture structure.

- The remainder of the architecture is the standard pipeline version, as discussed in Asudeh (2012), except that constituent-structure is now time-indexed. This indexation can be ignored for other purposes

- The gestural structure, combined in this way with the c-structure, contributes to the f-structure as a co-head of the projection of the node that directly dominates the gesture. The "syntactic" behavior of gestures can then be captured by the following rule:

  (1) $$\begin{array}{ccc} {}_iX_j & \rightarrow & {}_hG_k & {}_iX_j \\ & & \uparrow = \downarrow & \uparrow = \downarrow \end{array}$$

  where the time intervals $[i, j]$ and $[h, k]$ overlap, G is the category of gestural nodes and X is a metavariable for syntactic categories.

- 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 Correspondence Architecture, it can also be sensitive to information earlier in the pipeline, particularly f-structural information. Information extracted from f-structure can be used to appropriately instantiate the meaning of the gesture such that it takes into account morphosyntactic properties of its linguistic correlate.

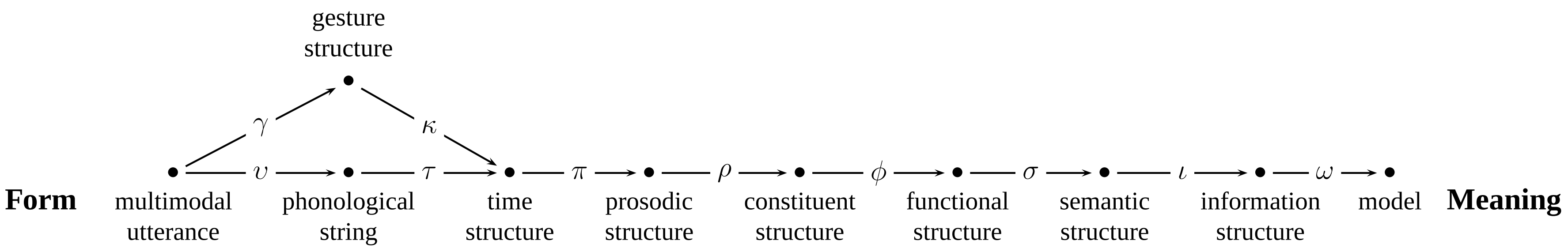


Figure 2: (Partial) Correspondence Architecture

6 Analysis

- To demonstrate the advantages offered by the projection architecture in modeling the integrated interpretation of gesture and speech, we will here briefly re-analyze an example presented in Giorgolo (2010), which is extracted from the Speech and Gesture Alignment Corpus (SaGA; Lücking et al. 2010), a corpus of spontaneous conversations annotated for gestural information. The example is shown in video 2.

- The speaker describes 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, namely that they are shaped like vertically oriented prisms and that the two towers are disconnected.

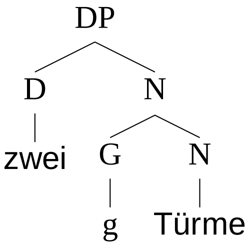
- From the raw, visual data (as represented in the video), the correspondence function γ generates a gesture structure for the multimodal utterance. A partial representation of the corresponding representation from the SAGA corpus is shown in (2).

(2)
$$\left[\begin{array}{ll}
\textsc{Left.HandShapeShape} & \text{loose C} \\
\textsc{Left.PathOfHandShape} & 0 \\
\textsc{Left.HSMovementDirection} & 0 \\
\textsc{Left.HandShapeMovementRepetition} & 0 \\
\vdots & \vdots \\
\textsc{Right.HandShapeShape} & \text{loose C} \\
\textsc{Right.PathOfHandShape} & 0 \\
\textsc{Right.HSMovementDirection} & 0 \\
\textsc{Right.HandShapeMovementRepetition} & 0 \\
\vdots & \vdots
\end{array}\right]$$

- The phonological string and the time structure together generate a time-indexed constituent structure in which the gesture is adjoined to the N node in the DP [zwei Türme]. This adjunction is defined in rule (3) and the resulting c-structure is shown in (4).

(3)
$$\begin{array}{cccc}
{}_iX_j & \rightarrow & {}_hG_k & {}_iX_j \\
 & & \uparrow = \downarrow & \uparrow = \downarrow
\end{array}$$

(4)



- The gesture and the noun it is adjoined are defined by rule (3) to map to the same f-structure. The gesture generally does not add f-structural information, but uses f-structural information in its f-structure to constrain interpretation and potentially places constraints on the f-structure it contributes to. The f-structure for (4) is shown in (5).

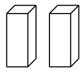
(5)
$$\begin{bmatrix} \textsc{pred} & \text{`tower'} \\ \textsc{number} & \textsc{pl} \\ \textsc{spec} & \begin{bmatrix} \textsc{pred} & \text{`two'} \end{bmatrix} \end{bmatrix}$$

- A partial general lexical entry is shown in (6) for a gesture that identifies two regions in space and here we show the semantic information for this gesture depending on whether it accompanies a two-place or one-place predicate. In this case, the two interpretations are determined by the NUMBER feature of g's f-structure, which is contributed by the plural noun 'Türme'.

(6) g G
$$\begin{array}{l} (\uparrow \textsc{number}) \neq \textsc{pl} \\ \lambda R.\lambda x.\lambda y.R(x,y) \wedge core(loc_e(x))(loc_e(y)) \\ ((\uparrow\ \textsc{obj})_\sigma \multimap (\uparrow\ \textsc{subj})_\sigma \multimap \uparrow_\sigma) \multimap \\ ((\uparrow\ \textsc{obj})_\sigma \multimap (\uparrow\ \textsc{subj})_\sigma \multimap \uparrow_\sigma) \\ \\ \qquad\qquad\qquad \bigvee \\ \\ (\uparrow \textsc{number}) =_c \textsc{pl} \\ \lambda P.\lambda x.P(x) \wedge (\delta(core))(loc_e(x)) \\ ((\uparrow_\sigma\ \textsc{var}) \multimap (\uparrow_\sigma\ \textsc{restr})) \multimap \\ ((\uparrow_\sigma\ \textsc{var}) \multimap (\uparrow_\sigma\ \textsc{restr})) \end{array}$$

- The function loc_e in the interpretations in (6) is a member of the family of Loc functions that map from the abstract frame of reference, F, to the spatial frame of reference, S (see Figure 1). The function loc_e maps entities to their spatial extensions.

- The two interpretations are distinguished by the number of arguments they take, but they are constructed around the same function, $core$, which represents the equivalence class of spaces that are informationally indistinguishable from the representational space. The function $core$ is defined in equation (7) for this example.

$core = \lambda r_1.r_2.\ (r_1 \cup r_2) \equiv$  (7)

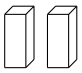
- The picture to the right of the equivalence in (7) is a short hand for a statement in description logic, part of which is shown in (8) for illustrative purposes.

$$\begin{aligned}
&\{\neg C(t_1, t_2), \qquad (8)\\
&\exists z.z(\forall w.C(w,z) \rightarrow C(w,t_1)) \wedge (\forall w.C(w,z) \rightarrow C(w,t_2 \oplus \mathbf{l})),\\
&\exists z.z(\forall w.C(w,z) \rightarrow C(w,t_2)) \wedge (\forall w.C(w,z) \rightarrow C(w,t_1 \oplus \mathbf{r})),\\
&\exists s.\exists z.\forall w.C(w, s \odot \mathbf{v} \oplus z) \rightarrow C(w,t_1) \wedge \neg(\exists z.\forall w.C(w, s \odot \mathbf{h} \oplus z) \rightarrow C(w,t_1)),\\
&\exists s.\exists z.\forall w.C(w, s \odot \mathbf{v} \oplus z) \rightarrow C(w,t_2) \wedge \neg(\exists z.\forall w.C(w, s \odot \mathbf{h} \oplus z) \rightarrow C(w,t_2)),\\
&\neg\exists z.z(\forall w.C(w,z) \rightarrow C(w,t_2)) \wedge (\forall w.C(w,z) \rightarrow C(w,t_1 \oplus \mathbf{u})),\\
&\neg\exists z.z(\forall w.C(w,z) \rightarrow C(w,t_1)) \wedge (\forall w.C(w,z) \rightarrow C(w,t_2 \oplus \mathbf{u})),\\
&\neg\exists z.z(\forall w.C(w,z) \rightarrow C(w,t_2)) \wedge (\forall w.C(w,z) \rightarrow C(w,t_1 \oplus \mathbf{d})),\\
&\neg\exists z.z(\forall w.C(w,z) \rightarrow C(w,t_1)) \wedge (\forall w.C(w,z) \rightarrow C(w,t_2 \oplus \mathbf{d})),\\
&\ldots\}
\end{aligned}$$

- In the information that corresponds to [NUMBER PL], we use a distributivity operator, δ, defined in (9). This operator decomposes the plural tower entity into singular towers (in this particular case, two towers, given the numeral), and their spatial projections are then restricted on the basis of the information conveyed by the gesture.

$$\delta(x) = \lambda e.x(e_1\ \cdots e_n) \qquad (9)$$

- The final interpretation is shown in (10). The interpretation corresponds to the characteristic function of properties that hold for two towers such that they are disconnected, vertical and square-based prisms. Part of the interpretation comes from the linguistic component of the multimodal utterance, 'zwei Türme'. The quantification and specification of the number of entities comes from the determiner 'zwei'. The predicate $tower$ comes from the noun 'Türme'. The rest is contributed by the gesture g. Finally, the bound predicate Q is the scope of the generalized quantifier ⟦zwei g Türme⟧.

$\lambda Q.\exists x.|x| = 2 \wedge tower(x)\ \wedge\ (\delta(\lambda r_1.r_2.\ (r_1 \cup r_2) \equiv$  $))(loc_e(x))\ \wedge\ Q(x)$ (10)

7 Conclusion

- The interaction between language and non-verbal communication is principled.
- The interaction is not just a matter of temporal alignment, but goes deeper, including interactions between non-verbal communicative elements and morphosyntactic information.
- In order to study this interaction, we need a rich and fine-grained framework, such as LFG's Correspondence Architecture.
- We have shown how to integrate non-verbal communicative elements in the Correspondence Architecture. In the case of hand gestures, we simply assume an additional gesture structure which is then integrated in the f-structure of the multimodal utterance.
- We have also shown how the Correspondence Architecture allows us to give a principled account of the influence of grammatical features like NUMBER on the interpretation of gestures. Similar effects can be envisaged for features like ASPECT in the case of gestures accompanying verbs or verb phrases.

References

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

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

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

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

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

Lücking, Andy, Kirsten Bergmann, Florian Hahn, Stephan Kopp, and Hannes Rieser. 2010. The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen, eds., *LREC 2010 Workshop: Multimodal Corpora Advances in Capturing, Coding and Analyzing Multimodality*.

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

Willems, Roel M., and Peter Hagoort. 2007. Neural Evidence for the Interplay Language, Gesture and Action: A Review. *Brain and Language* .