Multi-modal Integration for Gesture and Speech

Andy Lücking **Hannes Rieser** **Marc Staudacher**
Bielefeld University, CRC 360 "Situated Artifical Communicators", B3
{andy.luecking,hannes.rieser,marc.staudacher}@uni-bielefeld.de

Abstract

Demonstratives, in particular gestures that "only" accompany speech, are not a big issue in current theories of grammar. If we deal with gestures, fixing their function is one big problem, the other one is how to integrate the representations originating from different channels and, ultimately, how to determine their composite meanings. The growing interest in multi-modal settings, computer simulations, human-machine interfaces and VR-applications increases the need for theories of multimodal structures and events. In our workshop-contribution we focus on the integration of multimodal contents and investigate different approaches dealing with this problem such as Johnston et al. (1997) and Johnston (1998), Johnston and Bangalore (2000), Chierchia (1995), Asher (2005), and Rieser (2005).

1 Introduction

In this paper we are concerned with the multimodal integration of pointing gestures (called *gestures* hereafter) and speech. Gestures can be used to refer to objects present in the actual situation like apples or tables. It is also possible to point at objects not present in the actual situation as when giving directions or placing discourse referents into the gesture space (see McNeill, 1992). We confine ourselves to the former and provide crucial data for speech-gesture-integration below. We take these data as evidence for the claim that gestures are essentially linguistic.

A striking characteristic of the speech-gesture-interplay is that demonstratives (determiners, exophoric pronouns and place adverbs) *require* a gesture to co-occur with them. We represent a gesture's stroke with the symbol '↘', statements of acceptability are displayed as in (1) where '#' stands for "not acceptable".

(1) a. Grasp ↘ this bolt!
b. #Grasp this bolt!

Example (1-a) is well-formed while (1-b) is not, since the gesture is missing. In a related construction (replacing 'this' by 'the'), the use of a gesture is not required as the pair (2-a) and (2-b) shows.

(2) a. Grasp the bolt (on the table).
b. Grasp ↘ the bolt (on the table).

A feature left implicit in the format chosen to represent gestures and their co-present speech in (1) are the temporal relationships between them. Tokens of words and gesture can overlap in various ways. If we use a linear string representation of both words and a gesture's stroke with a precedence reading, different possible stroke positions give rise to different acceptability judgements. In other words, synchronisation matters. In case the stroke starts and ends before the onset of the accompanying utterance, as in (3-a), the multi-modal utterance has to be dismissed as being not acceptable. The same holds for strokes altogether following their affiliated speech, as in (3-d).[1] We conclude from these data that gestures have syntactic properties.

(3) a. #↘ Grasp this bolt!
b. Grasp ↘ this bolt!
c. Grasp this ↘ bolt!
d. #Grasp this bolt ↘!

Gestures also have semantic properties as the following example shows. Suppose a situation *s* where two candies, a red one and a green one, are lying side by side. Whether an utterance of 'This ↘ candy is red.' is evaluated as true or false in *s* depends on which candy is pointed at in *s*. Besides truth conditional effects, there is empirical evidence that gestures have *rich information content*. Lücking et al. (2004) found that the number of words used in a verbal description was less if the description was accompanied by a deictic gesture. Thus the finding suggests that gestures contribute content that otherwise would have to be expressed verbally.

Moreover, gestures relate to pragmatic phenom-

[1] However, we are able to interpret such utterances – presumably by pragmatic, i.e. inferential processes.

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 106–113.

ena. For example, it is not possible to substitute a verbal constituent for a deictic gesture in a *null context*, as in (4):[2]

(4) #He grasps ↘.

Note that example (4) can be rendered acceptable if a suitable object can be *accommodated*. Such a multi-modal utterance is also acceptable if it is uttered in a suitable context. For example, suppose a combat of gladiators in a Roman arena. The emperor decides whether they will live or die by pointing at them and (presumably) uttering (5-a) or (5-b), respectively. Given the supposed context, the utterances are acceptable.

(5) a. ↘ *missum*! (off he go!)
b. ↘ *iugula*! (cut his throat!)

In dialogues, a gesture can be used to realize a dialogue move. In (6) a piece of conversation between A and B is given, where B's gestural answer is acceptable. Its acceptability seems to be parasitic on the structure of question-answer-pairs and Gricean maxims.

(6) A: Where is the salt?
B: ↘

So, since gestures have syntactic, semantic, and pragmatic properties, they are just like words.

2 Interface Problems

If gestures are essentially linguistic, (formal) linguistic theories should account for them. From this point of view, current theories have a descriptive and explanatory gap and are in this sense deficient. Consequently, something new has to be taken account of. How shall we theorize? – In this section we discuss some *interface problems*.

The first point, however, relates to *theory change positions*. The question is whether a new kind of theory is required or an existing theory should be extended. Different answers are possible.

The *syntax enhancer* proposes a multi-modal theory not differing substantially form current ones. The enhancer thinks that syntax should be changed in such a way that gestures are accounted for, and then looks for changes in semantics and pragmatics.

[2]However, some readers might have different intuitions. We would like to point out that its acceptability might be due to the valence of the transitive verb predicting an argument at the level of its logical form which might be linked to ↘.

In opposition, the *syntax radical* proposes to develop a new kind of a multi-modal theory differing from the current ones in a substantial way. Properties of current theories need not be preserved. The radical thinks current syntax should be replaced by a new kind which can account for gestures from the outset and then looks for an apt semantics and pragmatics.

The *pragmasemantics enhancer* has the same attitude towards a multi-modal theory as the syntax enhancer has. However, he thinks that gestures should be accounted for in semantics and pragmatics as opposed to syntax. The enhancers seem to be more conservative than the radical.

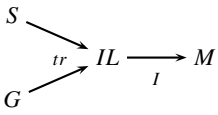
Each position has its price. By regarding gestures as linguistic, we change our existing concepts, notably some overly restrictive *concept of meaning*. Meaning is then no longer that which is or can be said but something else. A gesture cannot properly be "said". However, it seems that the richer concept of meaning still shares many properties with the traditional one.

The first problem relates to a consequence of the different positions, namely to the *point of integration*. The syntacticians propose either to approximate gestures to some existing linguistic category or to propose a new one for which combination rules are stated. The pragmasemanticists, on the other hand, will say that gestures are part of the linguistic context which is used to interpret an utterance. So, integration is some kind of syntactic combination (e.g. multi-modal subcategorization) and/or context-dependence.

The second problem relates to *descriptive adequacy*. There is an important difference between describing mono-modal information and multi-modal data. The former, e.g. speech, has a temporal order in such a way that for every two information bits one precedes the other. There are no overlap- or part-of-relations. In contrast, the information bits of the latter allow for such relations since the data is distributed across the different channels (such as sound and vision). Should an adequate description of multi-modal data take care of this? – This depends on the description's aim. For example, if an agent system is developed, multi-modal output planning might be important. Then questions of timing matter and, arguably, time should be explicitly represented. If the aim is doing semantics, however, only as much description is required as to describe the correct

satisfaction conditions. In this paper we take a linguistic view on the matter and have chosen a linear representation.

A related problem is the *linearization problem*. The question is whether all data descriptions have to linearize in the sense that the information bits in the representation have to be in a linear order. It seems to us that, when doing semantics in a type-logical-style, the data must always be linearized since every information bit in the representation can only be either a functor or an argument.The point is illustrated by the algebraic set-up of multi-modal integration below (tr is a translation mapping speech S and gesture G data to a type-logical intermediate-language IL. IL is interpreted by I in the semantics M.):



The fourth problem relates to *constructability*. It consists in providing a construction mechanism for logical forms of multi-modal utterances. If we have semantic aims, we want to have a systematic means to extract the right forms from multi-modal utterances. It should be possible to construct the intended logical form. Though, depending on the theory change position, what is needed can be quite different.

3 Approaches to integration

Having covered the ground for a review, we quickly chart out the proposals. A summary of the approaches is presented in Table 1.

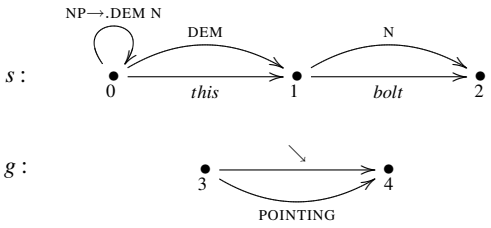
3.1 Johnston et al. HPSG (1997, 1998)

In course of the (military) software engineering project QuickSet, Johnston et al. (1997) developed an architecture that integrates input coming on different channels by means of unification of typed feature structures. The system provides a multi-modal interface allowing its user to give directives simultaneously by voice (speech) and pen (gesture) input. Both speech and gesture are assigned attribute-value matrix (AVM) representations by speech and gesture recognizers. Since a conventional HPSG grammar is merely extended to account for multi-modal utterances, it is an approach of a syntax enhancer.

Users interacting with the QuickSet system can point (at X) and by doing so they introduce a certain point in space represented as a latitude-longitude coordinate pair. This locational function of pointings is captured in the following representation showing that the semantics (content) of an object of category (cat) *spatial_gesture* is a definite point in space:

$$\begin{bmatrix} \text{cat}: & \textit{spatial\_gesture} \\ \text{content}: & \begin{bmatrix} \text{fsType}: \textit{point} \\ \text{coord}: \ \textit{latlong}(\text{x,y}) \end{bmatrix} \end{bmatrix}$$

The AVM-grammar formalism rests on a *multi-modal chart parser*. A multi-modal chart extends a conventional chart in that the former covers channel-crossing edges defined in terms of sets of identifiers of gestural (g) and speech (s) terminals:



Possible multicharts:

multichart 1: {[s,0,1], [g,3,4]}
multichart 2: {[s,1,2], [g,3,4]}
...

The basic rule allowing to "bridge" between the modalities is the *basic integration scheme*:

$$\begin{bmatrix} \text{lhs}: & \begin{bmatrix} \text{cat}: & \textit{comm} \\ \text{modality}: & \boxed{2} \\ \text{content}: & \boxed{1} \\ \text{time}: & \boxed{3} \end{bmatrix} \\ \text{rhs}: & \begin{bmatrix} \text{dtr1}: & \begin{bmatrix} \text{cat}: & \textit{loc\_comm} \\ \text{modality}: & \boxed{6} \\ \text{content}: & \boxed{1}[\text{loc } \boxed{5}] \\ \text{time}: & \boxed{7} \end{bmatrix} \\ \text{dtr2}: & \begin{bmatrix} \text{cat}: & \textit{spat\_gest} \\ \text{content}: & \boxed{5} \\ \text{modality}: & \boxed{9} \\ \text{time}: & \boxed{10} \end{bmatrix} \end{bmatrix} \\ \text{cnstr}: & \left\{ \begin{array}{l} \text{overlap}(\boxed{7},\boxed{10}) \vee \text{follow}(\boxed{7},\boxed{10},4) \\ \text{total-time}(\boxed{7},\boxed{10},\boxed{3}) \\ \text{assign-modality}(\boxed{6},\boxed{9},\boxed{2}) \end{array} \right\} \end{bmatrix}$$

The AVM for the integration scheme is stated very closely to a CFG-rule of the form lhs $\rightarrow$ rhs; the right-hand side (rhs) is made up of two constituents, namely dtr1 and dtr2. Thus, mapping the rule to a tree, they are the daughters of their mother constituent on the rule's left-hand side (lhs). The rhs-part of the AVM-structure is made up of a verbal located command (*loc_comm*; in QuickSet this can be, *e. g.*, "sandbag wall") and a spatial gesture. The gesture determines the location value

	Johnston HPSG et al. (1997, 1998)	**Johnston and Bangalore FSM (2000)**	**Chierchia (1997)**	**Asher SDRT (2005)**	**Rieser LTAG (2005)**
Motivation	Human Computer Interaction	Human Computer Interaction	Anaphora, context-dependent Quantifiers	Anaphora, anchoring of deictic NPs	Extended use of language, meaning, gestures as signs
Type of theory	Syntax, semantic representation	Syntax, semantic representation	Semantics, formal pragmatics	Semantics, formal pragmatics	Syntax, semantics, pragmatics
Type of grammar	Constraint-based (HPSG)	CFG	–	–	LTAG
Pointing representation	AVM-structure for locations	Object constants	Pragmatic indices	Externally anchored discourse referent	Set of object constants
Point of integration	Pointing introduced via subcategorization	Nouns, translation to semantics	Semantically underspecified quantifier representation	Presuppositional SDRS, underspecified discourse relation	Extended valence of relations in multi-modal interface
Strengths	Multi-modal chart parser	Highly efficient FSM parser	local extension of existing theory	Pointing in dialogue	Speech and gesture interaction in interface
Weaknesses	Restricted grammar, little linguistic motivation	No linguistic motivation, hardly extendible	No syntactic gesture representation	No syntactic gesture representation	No general mechanism to build interface from standard grammar

Table 1: Cluster of Approaches to Multi-modal Integration

of the word's content. The mother structure (lhs) then is a complete multi-modal command. The cross-channel integration is constrained by a set of restrictions given as the value of the feature cnstr (short for constraints). Most notably, *co-occurrence constraints* are expressed as temporal requirements, see the use of tags 7 and 10.

As it stands, the basic integration scheme licenses only multi-modal structures that consist of a speech portion and exactly one accompanying gesture. A more general framework that corrects this limitation is the extension given by Johnston (1998) where integration is handled via multi-modal subcategorization, analogous to the (lexicalist) treatment of complementation in HPSG. To this end, a new feature sbct (for subcat) is introduced whose elements can be recursively discharged by a *subcat combination scheme*, a generalized version of the basic integration scheme. Leaving the restricted QuickSet grammar but still remaining in the spirit of the grammar of Johnston and colleagues, the feature sbct can be used to capture that demonstratives require a co-occuring pointing gesture. A determiner like "this" is incomplete, that is, being of category *sub_dem*, unless it combines with the subcategorized gesture to build a proper AVM of category *dem*, as licensed by the subcat combination scheme. Applied to the example sentence "*Grasp this* ↘ *bolt!*", the "*this* ↘ "-part gets modelled as follows:

$$\begin{bmatrix} \text{lhs} : \begin{bmatrix} \text{cat} : & \textit{dem} \\ \text{cont} : & \boxed{3} \end{bmatrix} \\ \text{rhs} : \begin{bmatrix} \text{dtr1} : \begin{bmatrix} \text{cat} : & \textit{sub\_dem} \\ \text{cont} : \boxed{3} \begin{bmatrix} \text{obj} : \begin{bmatrix} \text{fsT} : \textit{exist\_there} \end{bmatrix} \\ \text{loc} : \begin{bmatrix} \text{fsT} : & \textit{point} \\ \text{coord} : \boxed{1} \end{bmatrix} \end{bmatrix} \\ \text{sbct} : \boxed{2} \begin{bmatrix} \text{cat} : & \textit{spatial\_gesture} \\ \text{cont} : \begin{bmatrix} \text{fsT} : & \textit{point} \\ \text{coord} : \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{dtr2} : \boxed{2} \begin{bmatrix} \text{cont} : \begin{bmatrix} \text{coord} : \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The demonstrative has a locational "gap" that gets filled by the gesture it subcategorizes for. The complete sentence is then projected as usual.

3.2 Johnston and Bangalore FSM (2000)

Johnston and Bangalore (2000) propose a multi-modal context-free grammar (CFG) to handle integration. Their parser implementation uses well-understood finite-state techniques.[3] Moreover, the translation to logical form is a product of concatenation. Thus, it is simple and highly efficient. As will become apparent, the position taken by this approach is the one of a syntax radical.

The multi-modal input, speech and gesture, is assumed to be distributed across different chan-

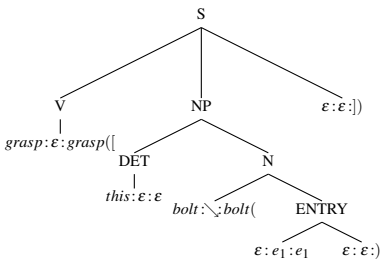
[3] Thus only the regular part of the CFG may be used.

nels. In order to use (mono-modal) context-free techniques, the reading of symbols in different channels is regarded as a single read operation of a complex, structured symbol. Each part of the structure relates to a symbol in a channel.

A *multi-modal CFG* is a tuple $\langle N, T, P, S \rangle$ where N is the set of *nonterminals* such as S, V or NP. T is the set of *terminals* of the form $W : G : M$ where W denotes a verbal symbol (e.g. 'bolt'), G denotes a gesture symbol (e.g. '↘') and M denotes a meaning expression, *e. g.* '*bolt*('.[4] All symbols of W, G and M are elements of the symbol alphabets Σ_W, Σ_G and Σ_M, respectively. Each alphabet contains the empty element ε. A special feature of the gesture alphabet Σ_G is that it consists of two disjunct subsets: One subset contains all gesture symbols, the other one contains all event symbols. As usual, P is the set of *productions* and S denotes the start symbol. Other notions such as *derivation* are the standard ones for CFG.

Gesture symbols and event symbols have different roles. An occurrence of the former indicates the presence of a gesture whereas the occurrence of the latter is used as a reference to entities referred to by a gesture. These event symbols label buffer elements of a finite buffer. The buffer is used to keep track of all assignments between gesture occurrences and the entities referred to by those gestures.

Using a modification of the sample grammar provided by Johnston and Bangalore (2000), the structure of (1-a) is:



The ENTRY node triggers the buffer mechanism which assigns e_1 a name of the entity referred to by the gesture occurrence, *e. g.* $obj1$. The meaning string $grasp([bolt(e_1)])$ is constructed by a top-down/left-to-right traversal through the tree. It is the result of the concatenation of the M-parts of every traversed terminal.

We think that their proposal is highly interesting to produce efficient parsers. However, it doesn't seem to be a good way to write *linguistically motivated grammars*. For example, sentences S also contain a verbal ε symbol. N is assigned a branching structure with ENTRY as its right node, and so on. Due to these *ad hoc* structures, the grammar is not easily extendible. The translation to logical forms is weird. There is no basic translation for 'this'. Basic translations for most syntactic expressions are not well-formed semantic expressions, *e. g.* '*bolt*(' or ')'. Thus it fails to be admissible on any standard account of semantic translation. For the same reason, an incremental interpretation is not possible.[5] Last but not least, using the "tape"-metaphor of automata theory, accumulation of gesture symbols on the tape leads to difficulties. Consider the unacceptable example (3-a). Since the gesture symbol is already on the tape, the N rule can be applied and thereby (3-a) is licensed by the grammar for (1-a).

3.3 Chierchia (1995)

Chierchia sketches a way to handle multi-modal integration in his renowned book *Dynamics of meaning* (Chierchia, 1995). He proposes to modify the translation of definites of the form *the N* to account for indexically used definites. Since he uses no syntactic representation for gestures and extends grammar conservatively, he is a pragmasemantics enhancer.

He locates the place where one should modify in the *representation of definites*.[6] In a first take he views a definite as a (partial) function from properties (represented as sentential functions) to the unique object that satisfies them, if there is such an object. He does so by introducing them formally as iota-terms of the form $\iota x \phi$ with the following semantics: If x is of type e and ϕ is of type t, then $[\![\iota x \phi]\!]^g = u$, where u is the unique object such that $[\![\phi]\!]^{g[x/u]}$, otherwise $\perp$ (read as "undefined").

However, a sentence like 'You, grasp the bolt.' being translated as $grasp(you, \iota x bolt(x))$ has an infelicitous use when the iota-term property is not satisfied. Chierchia's remedy is to analyse such utterances as utterances of the sentence 'You, grasp the bolt pointed at.' or 'You, grasp the bolt we are looking at.' which are translated as $grasp(you, \iota x(R(o, y, x) \wedge bolt(x)))$ and $grasp(you, \iota x(R(y, x) \wedge bolt(x)))$, respectively. In the translations o designates a location and y the speaker. The predicate R is interpreted as *is*

[4] The meaning expression is indeed written as `b-o-l-t-left parenthesis`.

[5] Arguably, it is possible using Lambda-terms.

[6] Note that the presentation of Chierchia's approach is simplified in that no possible world semantics is used.

pointed at ... by in the first translation and as *is looking at* in the second one.

Chierchia generalizes the translation of definites of the form *the N* to $\iota x(R(y_1, y_2, \ldots, y_n, x) \wedge N(x))$ where $R, y_1, y_2, \ldots, y_n$ are free variables. The context has to assign R an n-place function from the values of the pragmatic indices $y_1, y_2, \ldots, y_n$ to N's denotation. The indices $y_1, y_2, \ldots, y_n$ are taken to be part of the logical form.

So, the propositional part of (1-a) is analysed as $grasp(you, \iota x(R(o, y, x) \wedge bolt(x)))$ in a context where R, y and o are as before.

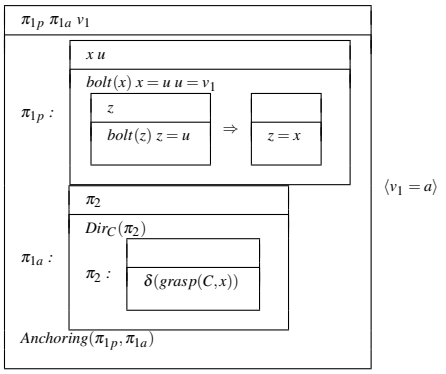
Chierchia's proposal is interesting, since it is a conservative extension of existing theories. The change is local and restricted to the logical form of definites. Though, does it amount to a satisfactory account of integration? We think not, since his proposal neglects the syntactic properties of gestures. They are not given a syntactic representation and only appear in the context. Moreover, there is no explicit integration mechanism. It is not clear how information given by gestures is used to construct the assignment for R, y and o.

3.4 Asher SDRT (2005)

Asher (2002, 2005)[7] sketches how SDRT can be extended to account for gestures. SDRT is a discourse representation theory modelling the semantics/pragmatics-interface. The theory itself is not committed to a particular grammar formalism and hence not to any specific syntax. Thus, the integration problem is approached by way of a pragmasemantic enhancer.

According to the SDRT account of definites, presuppositional information is distinguished from asserted information. The presupposed information of newly introduced definite NPs cannot simply be accommodated since an arbitrary object satisfying the conditions would not do for deictically used definites. It is proposed that definite descriptions introduce an underspecified relation, called *bridging relation*, between the referent and some other contextually given object, set to identity by default. In other words, such definites have to be *anchored* to some object in the non-linguistic context. Anchoring involves a *de re* attitude towards the object, some sort of *knowing how* needed to solve the conversational goals of the speaker. Anchoring requires linking an agent A's epistemic attitude to conversational goals. If an *Anchoring* relation between the presupposition of a definite ψ and some element in the discourse context exists for the agent A, he is supposed to have a computable means of getting to the referent of ψ from the present non-linguistic context of utterance under some given purpose ϕ; to capture this, a notion of *path* is defined. If the anchoring function of a deictically used definite is accepted by the participants in dialogue, they are assumed to mutually believe that the definite picks out the same object for them. Hence, anchoring amounts to coordination or alignment.

Applying Asher's new SDRT proposal to (1-a), the result of an apt multi-modal integration strategy under best-update is:



SDRT itself says nothing about multi-modal integration, though, it is part and parcel of it: The conceptual information of the gesture occurrence consists in the external anchoring of the discourse referent v_1 to the object a, written as $\langle v_1 = a \rangle$. The presupposed information of (1-a) is represented in π_{1p}, the asserted one in π_{1a}. The bridging relation between x and u is resolved to identity and thus $x = u$. We assume a speech act theory style imperative semantics. Consequently, Dir_C is to be read as 'C is commanded that ...' and $\delta(grasp(C,x))$ in π_2 is the action commanded, namely that agent C grasp x. Finally, the *Anchoring* relation holds between π_{1p} and π_{1a}. Thereby, x in π_2 is externally anchored to a.[8]

Asher's proposal is unique with regard to discourse modelling. However, gestures have no syntactic representation. It is not clear how the multimodal input is integrated. While the DRT construction algorithm can be used in principle, the construction problem remains unsolved in practice. For SDRT provides, "out of the box", neither

[7] We are grateful to Nicholas Asher for having taught us SDRT in the years 2003-2005 and for letting us work with unpublished SDRT material, especially Asher (2005).

[8] Observe that *Anchoring* is a subordinating discourse relation. Thereby, x in π_{1p} is accessible to x in π_2.

an anphora resolution mechanism nor a construction algorithm for SDRSs.

3.5 Rieser LTAG (2005)

In Rieser (2005)'s LTAG approach integration of demonstrations is handled by a grammar based interface. If, from the point of view of function, demonstrations are considered as words of a special kind, acting in a way like names, the lexicon of the interface has to be extended. It will encompass demonstration forms. In a similar way, syntax rules have to be added which allow for the combination of pointing and verbal expressions. Finally, gestural and verbal meanings have to be integrated in a compositional way. Here, interface modelling is based on LTAG (Joshi, 2004). This approach counts as a syntax radical. For verb valences in the interface are different from the usual ones. As a consequence, the denotation of verbs is also different.

Extension of LTAG Syntax We need additional structure in order to accommodate $\searrow$ and its positions. The LTAG-format used works with a set of trees anchored by terminal elements and two rules, substitution and adjunction. Adjunction will not be used, it is mentioned here for reasons of generality. $\searrow$ is considered as a terminal.

The relevant LTAG fragment is displayed in Fig. 1. (a) is the subject-less imperative rule. Elementary trees (b) and (b′) do service for two distinct pointing positions. They express that pointing is needed. (c) says that pointings function in a sense like NPs. (d) and (e) follow canonical CFG-rules.

Syntax-semantics Integration In order to achieve the syntax-semantics-integration we decorate all terminals with appropriate type-logical formulas. β-conversion is needed in order to model compositionality.[9]

$$\begin{aligned}
\text{grasp!}: &\quad \lambda\Theta\lambda\Pi\lambda u(\Pi(\Theta(\lambda y\lambda v F_{dir}(grasp(u,v)\\
&\quad \wedge(v=y)))))\\
\text{bolt}: &\quad \lambda x bolt(x)\\
\searrow: &\quad \lambda P_{\searrow}.P_{\searrow}(a),\ldots\\
\text{this}: &\quad \lambda P\lambda Q.Q(\iota x(P(x)))
\end{aligned}$$

How do we arrive at the representation of *grasp!*? – The reasoning is as follows: The verb *grasp* needs two argument slots, since it is transitive. In the interface fusing together the definite description and the pointing which goes with it, we need an identity condition linking the object argument of grasp and the variable for the object pointed at. Since the pointing functions like an NP, we must use the formula interpreting *grasp!* in order to get the correct bindings.

Computation of Meaning for (1-a) As can be seen from (1)-(8) the account is compositional:

$$\lambda\Theta\lambda\Pi\lambda u(\Pi(\Theta(\lambda y\lambda v F_{dir}(grasp(u,v)\wedge(v=y)))))\ \lambda P_{\searrow}.P_{\searrow}(a) \quad (1)$$

$$\lambda\Pi\lambda u(\Pi(\lambda P_{\searrow}.P_{\searrow}(a)\ (\lambda y\lambda v F_{dir}(grasp(u,v)\wedge(v=y))))) \quad (2)$$

$$\lambda\Pi\lambda u(\Pi(((\lambda y\lambda v F_{dir}(grasp(u,v)\wedge(v=y)))(a)))) \quad (3)$$

$$(\lambda\Pi\lambda u(\Pi((\lambda v F_{dir}(grasp(u,v)\wedge(v=a))))))\ (\lambda P\lambda Q.Q(\iota x(P(x)))\lambda x bolt(x)) \quad (4)$$

$$(\lambda\Pi\lambda u(\Pi((\lambda v F_{dir}(grasp(u,v)\wedge(v=a))))))\ (\lambda Q.Q(\iota x(bolt(x)))) \quad (5)$$

$$(\lambda u((\lambda Q.Q(\iota x(bolt(x))))\ (\lambda v F_{dir}(grasp(u,v)\wedge(v=a))))) \quad (6)$$

$$(\lambda u(((F_{dir}(grasp(u,(\iota x(bolt(x))))\ \wedge((\iota x(bolt(x)))=a)))) \quad (7)$$

$$F_{dir}(grasp(you,(\iota x(bolt(x))))\wedge((\iota x(bolt(x)))=a) \quad (8)$$

(8) is the result from β-reducing u in (7) with the indexical *you*.

4 Open Research Problems

The research on multi-modal integration (MMI) is still in its infancy. Therefore, basic empirical, methodological and theoretical issues have been hardly discussed. In this section we want to comment upon the following issues on the research agenda: Inverting the methodology for multimodal integration, motivation for a dynamic semantics approach, separation of presuppositional and assertional information, underspecification, restrictions of $\searrow$ under embedding, interaction of $\searrow$ with subsentential utterances, and interaction with iconic and emblematic gestures. These topics are treated in turn below.

As can be seen from the HPSG-approach and the LTAG proposal for MMI verbal expressions receive primary status on the modelling side and gesture is then added. One could also use the converse methodology, giving gesture primary relevance and adding language. It would drastically change the semantic role of gesture.

All the approaches discussed in this paper used static semantics, except Asher and Chierchia.

[9] In a way the strategy taken is similar to Lewis (1970) interpretation of PSG with formulas of intensional logic.

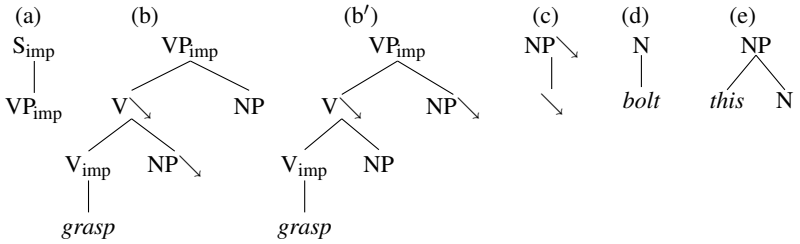


Figure 1: LTAG Fragment

However, there are good reasons for a dynamic semantics account. Such a dynamic account can also be fruitfully applied to complex demonstrations functioning as antecedents (7-a) or as anaphora (7-b):

(7) a. You may have ↘ this piece of cake. It tastes awful.
b. You are looking for a sweet? Take ↘ this strawberry tart.

What one can learn especially from the SDRT account is that presuppositional and assertional information should be separated. As a consequence, the use of rhetorical relations seems mandatory.

As example (3) shows, stroke positions can appear at various places in the utterance, stroke is, to borrow a linguistic term, polymorphic. A detailed reconstruction of this effect in grammars would have to result in a plethora of rules. Therefore an underspecification account of "stroke-syntax" seems to be more advisable. A similar argument goes for the pairing of stroke positions and functions of stroke. If position determines function, as seen from a type-logical perspective, we have many functions, which, however, are perhaps not distinct from the semantic point of view. This is hard to model.

Empirical and modelling problems arise, if strokes appear with deeply embedded material. This may give rise to ambiguities concerning attributions of stroke which is relevant with regard to truth conditional considerations. In the situated communication data used here, pointing comes frequently with subsentential utterances. A study of this effect using SDRT as the descriptive frame was started in (Lücking et al., 2006).

We know from our experimental studies that pointings tend towards "iconization". It is not clear as yet, how these effects should be modelled. One interesting aspect is how to represent iconic gestures and how to deal with compositionality matters in the interface of demonstration and iconicity.

Acknowledgments

Our work on SDRT was supported by the CRC "Situated Artificial Communicators," project "Deixis in Construction Dialogue" (DEIKON) at Bielefeld University, funded by the German Research Foundation (DFG). Thanks to the anonymous reviewers whose remarks were helpful for improving our paper.

References

Nicholas Asher. Deixis, Binding and Presupposition. Forthcoming in: Festschrift for Hans Kamp, 2002.

Nicholas Asher. Bielefeld Lectures on SDRT, 2005.

Gennaro Chierchia. *Dynamics of meaning: anaphora, presupposition and the theory of grammar*. University of Chicago Press, Chicago [a.o.], 1995.

Michael Johnston. Unification-based Multimodal Parsing. In *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics – Volume I*, pages 624–630, Montreal, Quebec, August 1998. ACL.

Michael Johnston and Srinivas Bangalore. Finite-state Multimodal Parsing and Understanding. In *Proceedings of the 18th Conference on Computational Linguistics – Volume I*, pages 369–375, Saarbrücken, July 2000. ACL.

Michael Johnston, Philip R. Cohen, David McGee, Sharon L. Oviatt, James A. Pittman, and Ira Smith. Unification-based Multimodal Integration. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pages 281–288, Madrid, July 1997. ACL.

Aravind K. Joshi. Starting with complex primitives pays off: Complicate locally, simplify globally. *Cognitive Science*, 28(5):637–669, 2004.

David K. Lewis. General semantics. *Synthese*, 22:18–67, 1970.

Andy Lücking, Hannes Rieser, and Jens Stegmann. Statistical Support for the Study of Structures in Multi-Modal Dialogue. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog '04*, pages 56–63, Barcelona, 2004.

Andy Lücking, Hannes Rieser, and Marc Staudacher. SDRT and Multi-modal Situated Communication. ESSLLI, to appear, 2006.

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

Hannes Rieser. Pointing and grasping in concert. In Manfred Stede, Christian Chiarcos, Michael Grabski, and Luuk Lagerwerf, editors, *Salience in discourse: multidisciplinary approaches to discourse*, pages 129–139. Nodus Publikationen, Münster, 2005.