Prerequisites for a Multimodal Semantics of Gesture and Prosody

Dafydd Gibbon

2004–12–15 (IWCS-6 2005)

1 Basic assumptions

Prosody and gesture in speech, text and images in documents, even music and other art forms have been included in the scope of multimodal communication. In its most general sense, multimodal communication may thus include any combination of sequential and parallel human motor and sensory facilities used in any kind of communication. The present contribution is more restricted, and covers gesture, with supplementary reference to prosody. Gesture is taken to be

not the highly articulated sign languages used in acoustically hostile situations, such as the gestural languages of the deaf, of communities with speech taboos, of nautical or aircraft ground control semaphore, or of the stock exchange,

but rather conversational gesture involving movements of head, limbs and body, including both highly conventionalised *emblems* and the more highly situated *gesticulations*.

In the present approach[1] (see also [7]) the specific claims are made that conversational gestures are

1. speech–like,
2. in particular, prosody–like (a frequently noted similarity),
3. articulated sequentially or in parallel,
4. compositional, being grounded in morphemes and words in a gesture lexicon [17].

A useful analogy is provided by Browman & Goldstein in Articulatory Phonology [3]. They have convincingly shown that the speech production is best explained in terms of task dynamics applied to a theory of physical gestures. The gestures of speech are, in this approach, no different from other gestures in terms of their structure and their physical realisation, including their temporal properties, except that they transduce muscle effort into sound (indeed there are several other kinds of gesture which also do this).

A fruitful strategy of investigation of conversational gesture therefore does

[1]Thanks are due to members of the DFG–funded ModeLex project at Universität Bielefeld particularly to Alex Thies, for innumerable discussions, and for their contributions to joint development of the CoGesT gesture notation.

not involve intuitive and ad hoc terminology specific to the domain (as was often the case in prominent studies during the past 30 years or so), does

not use space coordinate matrices and vectors of quantitative positions, angles and speeds, as in robotics–based gesture study,

but rather employs categories which are analogous to the semiotic categories of speech.

Informal categories such as those introduced, and convincingly motivated, by McNeill in many publications (e.g. [15]) will be used as starting points for a discussion of the relevance of gesture forms and functions to the conceptual systems commonly used in computational semantics, and as anchors for linking gestural systems with spoken language. These categories (which may be better seen as parameters) include *emblems* (highly lexicalised, and not 'gesticulations' proper), *iconics* ('imagistic' gestures involving similarities of shape, direction, speed or angle between a gesture and its referent), *metaphorics* ('imagistic' gestures involving a vehicle or base image which the gesture depicts, and a tenor, referent or abstract meaning of the metaphor), *deictics* (indexical, pointing gestures of direction or position), *beats* or *batons* (regularly repeated gestural movements which accompany speech accentuation), *Butterworths* (gestures which accompany speech disfluencies).

However, several of the points listed here are contrary to the basic tenets of McNeill's approach, in which gesture is

not morphemic,

not compositional,

not defined in terms of form–meaning conventions,

but rather global, in that the meanings of the parts are determined by the meaning of the whole in which they are embedded.

The strategy which is pursued in the following sections is explicitly linguistic, using well–tried linguistic constructs as analogies for describing gestural forms and functions, and using existing informal accounts of gesture as explicanda for this purpose.

2 The submodality of prosody: metalocutions

A traditional view of parallel information streams in face–to–face communication relates to prosody, and prosody will serve as a useful introduction to the topic of multimodal semantics. Intonation has often been seen as a concomitant parallel channel of communication. This concomitant signalling system has sometimes been seen as subordinate to the verbal or locutionary component of speech, for instance as a convenient focus disambiguator or as a source of speech act markers, and sometimes as a source of additional meanings which are not otherwise conveyed by the verbal or locutionary component of speech.

In [5] the term *metalocutionary* was coined for one functionality of parallel prosodic channels in relation to a basic *locutionary* channel of information. The

idea behind this innovation was that the parallel data streams of prosody and the lexico–syntactic components of locutions are in a *semantic* relation to each other. The approach rides rough–shod over the carefully tended lawns of formal semantics, but the basic idea is clear: prosodic forms and constructions are defined in a 'prosodic language' which has

1. its own syntax, defining sequences of boundary tones, accents, and concurrent longer and slower contours,

2. its own denotational semantics, in which components of the intonational syntax are mapped to a domain of locutionary constituents which are temporally co–extensive (synchronised, parallel) and — by virtue of being uttered by the same person — also spatially co–extensive.

In [6] the term *metadeixis* was used for specific metalocutionary functionalities in which terms in the 'language' of prosody literally *refers* to co–extensive units of language: an accent, for example, *denotes* the constituent it is associated with in the same way in which a pointing gesture denotes the object to which it points.

This approach is pre–figured in traditional functional descriptions of prosody, in which terms such as *configurational function*, with complementary components of *delimitative function* and *culminative function*, are widely used. Perhaps the most well–known functions of this kind are associated with the 'onset', 'nuclear tone', 'tail' and 'paratone' categories of functionalist [10] and discourse theoretic [9] intonation descriptions, and with the 'boundary tones' of current level–tone based notations such as ToBI transcription [20].

In the following sections of this contribution, a closely related approach wil be proposed for other information channels in multimodal communication, with the suggestion that this semantic approach to parallel information streams, in which components of parallel streams may denote components of other parallel streams, offers advantages over a purely syntactic approach to event overlap and precedence.

First a basic linguistically motivated sign model will be outlined, and terminology to do with multimodality will be clarified. Then semantic, pragmatic and syntactic explicanda for multimodal sign complexes will be discussed. Finally, a first attempt at capturing the syntax of brachial and manual gestures and their semantics will be outlined.

The claims are based on extensive empirical work, but this contribution does not aim to provide the empirical evidence. The main goal is to provide a set of explicanda for integrating gesture descriptions into current methodologies in formal and computational semantics, from the point of view of an ordinary working linguist who uses computational methods for checking formal consistency, for corpus analysis, and for creating practical applications.

3 Signs, modalities and submodalities

The sign structure presupposed here has

1. a core of structural *units* related by a *syntax* which defines occurrence in *external contexts* and *internal composition*, and

2. a *semantic interpretation* of the units on the one hand, and a *modality interpretation* (analogous to phonetic interpretation) on the other.

This kind of structure is common in contemporary linguistic theories, such as HPSG and those of the MIT school. The approach differs from standard approaches in logic, for which a component of modality interpretation (or phonetic interpretation) is alien, the form of signs being in general identical with their spelling.

Multimodal signs have been studied for many decades, mainly on the periphery of established disciplines like linguistics, psychology, and rhetoric ('body language'). The popularity of multimodal domains during the past half–decade or so is due largely to their increasing importance in natural language communications and assistive technology engineering, in which speech is now seen as multimodal and not just acoustic. Similarly, the increased prevalence of mixed modality documents in various internet services — text, images, video, audio — has led to a view of text as something much more complex than simply a stream of language units with media–specific layout.

The following system–oriented definitions are taken from [8] p. 105:

Multimodal system: a system which represents and manipulates information from different human communication channels at multiple levels of abstraction.

Multimedia systems: a system which offers more than one device for user input to the system and for system feedback to the user.

The times are a–changing, though, and these definitions do not do justice to the current situation. Therefore I will use the following definitions, based on a simple and fairly standard model of interpersonal communication via a channel:

Medium: A visual or acoustic channel of communication which may be *natural* (e.g. within the ranges of hearing or vision) or *technical* (e.g. supported by sensor, amplifier, recording, transmission, and display artefacts). Face–to–face speech and gesture classify as natural media, telephone speech as a technical medium. Writing classifies as a technical medium.

Modality: A pair of a human output device, which modulates a signal in a visual or acoustic channel, and a human input device, which de–modulates this signal. For present purposes, human output devices are the voice, the head and the limbs, and the input devices are the eye and the ear.

Submodality: An independent or near–independent modulation in a given channel by a sub–component of a human output device, demodulated by a separate component of the input device. Intonation, based on a vibration generated by air pressure and vocal cord tension, and locutionary patterns, generated by other obstructions of the vocal tract, are submodalities of the vocal modality. Submodalities of written documents are text and images.

The senses of touch, scent, and taste can also be included in the definitions; they are not so relevant to the present discussion, however. These definitions permit interesting variants on traditional definitions. For example, vocal gestures are not the only gestures which transduce muscular energy into sound; clapping, finger–snapping, foot–stamping, lip–smacking and whistling are also of this type.

4 Pragmatic explicanda

In [2], Allwood follows a strategy of taking well–established categories of a communication model (sender, recipient, expressions, media, content, purpose, environment), and a semiotic model (index, icon, and symbol), among others, as explicanda for use in classifying the functions of gesture. A similar strategy is pursued here, but distinguishing between well–known pragmatic functions (dialogue management, speech–acts, turntaking, backchannelling) specifically semantic functions (predication, naming, quantification, metaphor).

The idea of prosodic forms as markers of dialogue functions is very old. Traditional, over–simplified functional terminology such as 'question intonation' begs the question of whether there are simple relations between intonational forms and functions, but it does express the basic intuition that prosody has functions in dialogue.

The specific idea that intonations are 'illocutionary force markers' goes back to Austin and Searle, and was taken up by explicitly in linguistics by [13]. The idea was partially refuted in [5] by demonstrating that in the clearly identifiable case of call contours, the conditions for use are not illocutionary (in the sense of Searle [19]) but rather the infringement of Searle's first condition that normal input and output conditions obtain, i.e. a pre–condition for any communication, not the specifically illocutionary essential condition. The illocutionary hypothesis could be rescued for these patterns, of course, by referring to input/output conditions in the essential condition. But still, there a clear difference between this kind of metalocutionary illocution and the classical kinds. In traditional terms, these illocutions could be referred to as 'phatics'. The same call contours tend to be used in other phatic contexts, such as chanted openings and closings — "Hello–o!", "By–ye!", "Thank–you!". In German (cf. [5]) there is an additional use for these contours when communication breaks down owing to overly quiet or indistinct speech — "Lau–ter!" — or in repetitions: "Nein!" — "Wieso?" — "Ne—ein!" (or even 'Na—jen!" in the Bielefeld area dialect).

These phatic functions can also be observed in gestural modalities, for example *waving* in order to attract attention or to greet, and *hand–shaking* in order to establish contact, literally, either when opening a channel of communication or closing one, or cupping the hand around the ear if the signal is to quiet or too noisy.

A clear case of a speech act function is found in gestures which have the perlocutionary function of insulting, possibly in addition to other functions such as rejection. Examples of these are obscene gestures such as 'the finger' or the palm–up vee-sign, or complaining gestures such as tapping the side of the forehead (known in German as 'jemandem einen Vogel zeigen'). Such gestures are highly stylised emblems; the repertory of such gestures may be seen as subsets of a gestural lexicon of easy to identify items. Less clear as speech act gestures are such cases as as the thumbs up or thumb-index-circle 'ok' gesture, roughly meaning agreement or fulfilment of expectations.

Speech act gestures are closely connected to turn–taking devices. It has been well–known since the 1970s that turn–yielding, turn–defending and turn–claiming signals include gestural, involving posture, arm–movements (e.g. gesture–retraction) and gaze.

Adjacency tuples such as question–answer–confirmation tend to be struc-

turally marked by both prosody and gestural patterns. The following[2] are common:

- raised eyebrows: surprise, back–channel query.
- shoulder–shrugging: a response, paraphrased by 'I don't know'.
- head–nodding: agreement.
- head–shaking: disagreement.
- head–rocking (from side to side): neither yes nor no; undecided.

It has been observed that in cooperative contexts, speech patterns at all levels — prosody, vocabulary, syntax — are adapted to the prosody of the interlocutor. This happens most clearly in 'motherese' and 'lovers' talk': in talking to a baby or small child, or an intimate partner, pitch height, rhythm and tempo may be adapted (not a necessity), and vocabulary restrictions and grammatical simplifications may be used. Similarly, in professional cooperative bi–gender communication pitch raising by men and pitch lowering by women, as well as other forms of group adaptation and social alignment, can also be observed.

Phatic functionality in cooperative environments demonstrates consensuality (cf. [1]), and is an important factor in cooperative systems design. Phatic functionality is also observable with posture, in particular with head and arm positioning. For example, photographs of a mixed group of people will often show some members with roughly the same posture. I term this *postural harmony*: the closer their social relationship, the more similar the postures are likely to be.

Ludic gestures and emblems are used in interactive games, and and, finally, 'magical' gestures such as knocking on wood, finger–crossing (British) or thumb–squeezing 'Daumendrücken' (German) are intended to bring good luck.

4.1 Semantic explicanda

A gesture, and in particular an emblem, may often be interpreted as an elliptical *predication* in which the arguments of the predicate are the speaker and the addressee. Tapping the side of the forehead, for example, may be paraphrased as: "I am communicating to you that you are an idiot." A hand–wave may be paraphrased as "I have seen you and assure you that commuinication between us is okay." A thumbs–up sign may be paraphrased as "I communicate a positive judgment." Conversely, the Roman "thumbs down" sign may be paraphrased as "I communicate a negative judgment." An alternative analysis would be to regard a gesture as a complete proposition; no position will be taken on this at present.

Gestures may be used for *naming*, most clearly in deictic gestures, specifically, in pointing. In fact, in a sequence such as the following

A: finger points at B, hand waves towards door, ending up pointing to door.

[2]Note that here, as elsewhere in this discussion, form–function relations of gesture, whether emblems or gesticulations, are highly language and culture specific.

B: leaves through door.

a gestural utterance involving definite descriptions is constructed which can be paraphrased very crudely as

x = B & y = the–door & AGENT(x) & ROUTE(y) & LEAVE–BY(x,y)

Gestural *quantification* of different ontological objects is also found:

1. Counting gestures (with both cardinal and ordinal function),
2. part–whole gestures,
3. size gestures,
4. speed gestures,
5. direction gestures.

Gestural *conjunction* can be observed in more complex contexts, for instance marking "on the one hand .... on the other hand" types of argumentative construction.

In the discussions of *metaphorical*, *ironic*, and *double–bind* functions of gestures, the underlying common feature is that, like lying, all involve different kinds of discrepancy between literal message interpretation and reality.

McNeill [15] distinguishes between iconic gestures and metaphors, but using conventional criteria this distinction may be resolved into a single overall category of metaphor (to some extent this also occurs in McNeill's more recent work, largely based on Lakoff's conceptual model approach). So, venturing into the lions' den, and combining elements of the definitions of Aristotle, Max Black, John Searle and George Lakoff, metaphors will be defined as follows for present purposes:

Metaphor: A kind of discrepancy between message interpretation and perceived reality, in which the interpretation of the utterance is constructed on the basis of similarities between a conceptual model of the tenor *comparandum* and a conceptual model of the vehicle *comparatum*, and the assignment of other (typically appraisive) characteristics of the vehicle to the tenor.

Many stylised, lexicalised gestures are *iconic*, in that they have similarities of shape, trajectory or extent to the objects or events to which they relate semantically. A well–known example is the telephone gesture, with a basic fist configuration, but with thumb and pinky extended. Iconics are a special case of metaphors, in that they tend not to transfer an appraisive judgment from the vehicle to the tenor of the metaphor, but only utilise similarity criteria.

There are also typographic and prosodic icons with a similar basic pattern. A well–known exaple of a typographic icon is Lewis Carroll's "tale of a tail", in which the story about a tail is laid out visually to look like a stylised tail. Or in a children's story about a tiny little mouse and a big brown bear the following iconic typography might be used:

> Once upon a time there was a $^{\text{tiny little mouse}}$ who lived together with a **BIG BROWN BEAR**.

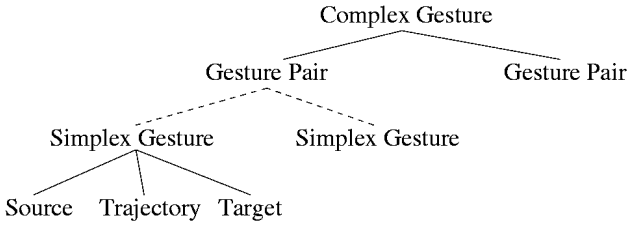

Figure 1: Gesture morphology.

Reading this to a three–year–old relative one could pronounce "tiny little mouse" with a high pitch and fast tempo, and "big brown bear" with a low pitch and slow tempo; the opposite pronunciations would be incongruous (though funny).

Gestures may also be used to mark irony, i.e. another kind of discrepancy between message interpretation and reality, and it has often been noted that double–bind communication involves a verbal message and a gestural message which convey incompatible meanings.

The metalocutionary hypothesis of gestures being used like prosody was discussed in the introductory sections. Gestures may be used in metalocutionary functions — cf. McNeill's category of 'beat' or 'baton', for example, as marking foci in speech in the same way as accentuation.

5 Syntax of CoGesT

The CoGesT notation and annotation system was developed in order to provide a first approximation of a lingusitically and computationally useful gesture transcription scheme. The application domain of CoGesT is currently arm and hand gestures.

The fundamental object described by CoGesT transcriptions is the *Simplex Gesture*. The Simplex Gesture has an obligatory source specification (the location of the hand or arm in space) and an optional route specification (the movement of the hand or arm in space). Gestures which consist only of a source are static gestures such as postures and held movements (or holds). Gestures which have a source and a route are *dynamic gestures*, and include a movement. The route consists of a trajectory and a target. A dynamic gesture is consequently fully specified by its source (the starting point), the target (the end point) and the trajectory between these two points. In McNeill's terminology the starting point would be at the onset of the preparation phase, the movement would be the process of transition between preparation phase via stroke back to the retraction phrase, of which the final position would be the target.

5.1 Data structures for Simplex Gestures

The basic data structure used to transcribe the Simplex Gesture is a feature vector. The visual semantics of Simplex Gestures defines postures or movements which are carried out with one body part or limb only. Simplex Gestures are only compositional with respect to their internal 'gestalt', somewhat like complex phonetic units such as stop consonants. The notion of Simplex Gesture abstracts

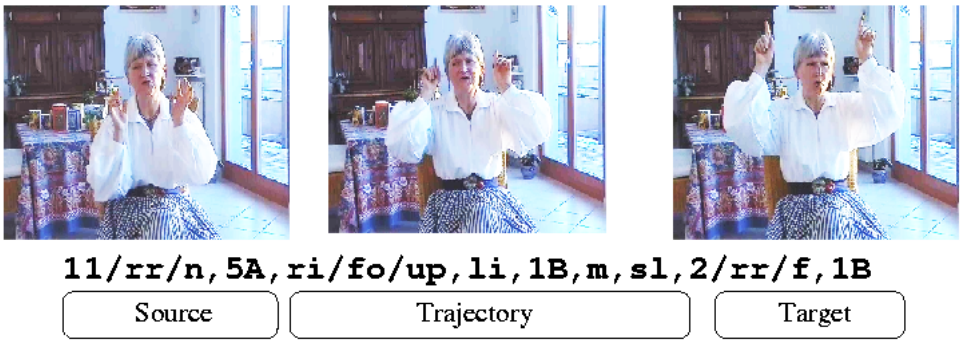

Figure 2: Video narrative, with CoGesT vector transcription.

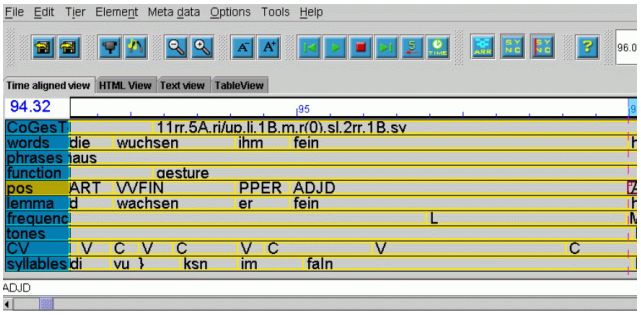

Figure 3: Multi–tier annotation using the TASX–annotator

away from functional categories, from concatenations of gesture sequences, and from associations of simultaneous movements of different limbs.

Figure 2 illustrates the transcription of the right hand part of a gesture that is performed with both hands, starting with relaxed hands at neck height and moving upwards with pointing hands. This example is taken from a corpus where at the same time the growing ears of a donkey are described on the spoken tier.

Figures 2 and 3 show the gesture transcription embedded in a multi–tier annotation of the video recording.

A CoGesT transcription vector, as shown in Figure 3, consists of:

1. The location specification for gestures, which refers to a virtual grid over the space in which a body is located. This grid is not meant to be absolute but relative to one's perception, specifying a perceived location in respect to horizontal (19 horizontal divisions), vertical and sagittal (5 divisions each) planes.

2. The shape of the hand, which is currently described iconically by 48 different prototypes that correspond to the handforms used by [18] and [14].

3. The movement (if any), which is described in terms of

   - the direction of a movement, which is given in a vector for all three axis relative to the previous location,

Figure 4: Virtual grid for location specification

- the shape of the movement, which is described in 7 elementary time functions; for more complex movements the shape of the movement is expressed as an iterative time function with iterations referred to as *microgestures*,
- the shape of the hand during the movement,
- a description of the size of a gesture and the speed of the movement,
- the target location.

For practical applications the fuzziness of this method is accepted in order to allow integration into a multi–tier score with all sorts of other annotation levels, such as prosodic or orthographic annotation or glossing. Figure 3 illustrates this kind of multidimensional annotation.

The CoGesT vectors could easily be described by a Regular Grammar or Finite State Automaton: any hierarchical grouping they may be given has a finite depth, and any recursion they may have is iteration, i.e. tail recursion. However, for use in potential semantic interpretations it seems advisable to think at least in terms of a Context–Free Grammar. For this reason, a context–free grammar in EBNF notation was defined (and is in fact used in a verification parser for annotation input):

```
<cogest>              ::= <complexgesture>
<complexgesture>      ::= <gesturepair>[<complexgesture>]
<gesturepair>         ::= <simplexgesture><simplexgesture>
<simplexgesture>      ::= <source>[<route>]
<source>              ::= <location><handshape>
<route>               ::= <direction> (<trajectoryshape> | <microgesture>)
                          <trajectoryhandshape> <trajectorysize>
                          <trajectoryspeed><target>
<microgesture>        ::= <source><route>[<microgesture>]
<direction>           ::= <lateral><sagittal><vertical>
<lateral>             ::= ri | le | NULL | ?
<sagittal>            ::= fo | ba | NULL | ?
<vertical>            ::= up | do | NULL | ?
<trajectoryshape>     ::= ci | li | wl | ar | zl | el | sq | ?
<trajectoryhandshape> ::= <handshape>
```

```
<trajectorysize>      ::= xs | s | m | l | xl | ?
<trajectoryspeed>     ::= sl | fa | me | ?
<target>              ::= <location><handshape>
<location>            ::= <height><verticalpos>
<height>              ::= 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
                          13 | 14 | 15 | 16 | 17 | 18 | 19 | ?
<verticalpos>         ::= ll | l | m | r | rr | ?
<handshape>           ::= 0A | 1A | 2A | 3A | 4A | 5A | 6A | 0B | 1B | 2B |
                          3B | 5B | 6B | 0C | 1C | 2C | 3C | 5C | 6C | 0D |
                          1D | 2D | 3D | 5D | 6D | 0E | 1E | 2E | 3E | 5E |
                          6E | 0F | 1F | 2F | 3F | 5F | 6F | 1G | 2G | 5G |
                          6G | 5H | 6H | 2I | 5I | 6I | 2J | 2K | 7A | ?
```

The values height and vertical position terms refer to the body grid, and the handshape term refers to the shapes defined in [14].

More conventionally, by contemporary standards, the CoGesT representation can be thought of as a convenient abbreviation for a more elaborate attribute–matrix (AVM) notation. The following AVM represents the categories encoded in the transcription in abbreviated form; the AVM shown here is embedded into a more complex gestural sign structure:

$$\begin{bmatrix}
\text{Source:} & \begin{bmatrix}\text{horizontal_plane:} & \text{neck height (11)}\\ \text{vertical_plane:} & \text{right_of_body (rr)}\\ \text{sagittal_plane:} & \text{central position (n)}\\ \text{handshape:} & \text{neutral, relaxed hand (5A)}\end{bmatrix}\\
\text{Trajectory:} & \begin{bmatrix}\text{direction:} & \begin{bmatrix}\text{horizhontal_axis:} & \text{right (ri)}\\ \text{sagittal_axis:} & \text{forward (fo)}\\ \text{vertical_axis:} & \text{upward (up)}\end{bmatrix}\\ \text{trajectoryshape:} & \text{straight_line (li)}\\ \text{trajectoryhandshape:} & \text{extended_index_finger_others_fist (1B)}\\ \text{trajectorysize:} & \text{medium (m)}\\ \text{trajectoryspeed:} & \text{slow (sl)}\end{bmatrix}\\
\text{Target:} & \begin{bmatrix}\text{horizontal_plane:} & \text{top_of_head (2)}\\ \text{vertical_plane:} & \text{right_of_body (rr)}\\ \text{sagittal_plain:} & \text{front_position (f)}\\ \text{handshape:} & \text{extended_index_finger_others_fist (1B)}\end{bmatrix}
\end{bmatrix}$$

5.2 Conclusion: linguistic accounts of multimodality

It is not too hard to relate prosody to the locutionary forms of language, at all levels:

1. In lexical tone languages, the tone functions exactly like a phoneme. In tone languages with both lexical and morphosyntactic tones, like the Niger–Congo languages of Western and Central Africa, lexical tone functions exactly like a phoneme, and morphosyntactic tone functions exactly like a morpheme: a high tone may indicate future tense, a low tone present tense.
2. In intonation languages, the tone may also function like a morpheme:
   - some languages have interrogative particles, some have marked word order, some have rising tones, some have various combinations of these;
   - some languages have focus particles, some have marked word order, some have specific pitch accents, some have various combinations of these.
3. It appears that all languages there is a concept of *prosodic domain*, over which interaction within sequences of prosodic units is observed, and which is at least

partially determined by phrasal syntax: in languages like English, patterns of the Nuclear Stress Rule type are observable, though, unlike other grammatical rules, the rule only defines a tendency and many other factors are involved. It has been shown that Noun Phrases in Niger–Congo languages, for example, determine the domain of tone terracing, a form of tone sandhi.

For gestural modalities the task is harder. The classic distinction between *preparation*, *stroke* and *retraction* is exactly comparable to the phases of gestural phonology, and there are structural analogies with theories of higher–level units such as the syllable, structured into onset, peak and coda. Perhaps a comparison with sequences of unstressed, stressed and unstressed syllables which characterise word and phrase structure is legitimate, at least for beats and batons, perhaps also for deictics. This kind of comparison is of course so general as to be of little theoretical value. Nevertheless, it does provide a structural analogue as a heuristic justification for using the same notations in preparation for further integration of descriptions of gestural patterns into other linguistic patterns, rather than *ad hoc* notations. However, at least the idea of prosodic domains is informally present in the notations used in [15]

It would be premature at this point to attempt a formal explication of the pragmatic and semantic explicanda noted in previous sections; in any case, this is not the main aim of this contribution. However, the discussion there provides several of pointers as to how the semantics of gesture could be integrated into the semantics of vocal utterances. But the metalocutionary hypothesis (cf. [5], [6]) noted in previous sections for prosody may be spelled out in a little more detail as follows:

1. Pitch accents *denote* the focus operands in the locution.
2. Final pitch accents (nuclear tones) *denote* the main thematic focus in the locution.
3. Boundary contours such as final lowering, final raising, initial onset tones *denote* the boundaries of units in utterances.
4. Slow–moving pitch contours between initial and final boundary tones *denote* the extents of units in utterances.
5. Specific pitch patterns *denote* functional components of speech acts, e.g. (subject to much dialectal and stylistic variation):
    - rising tones *denote* a non–final element of a series of units in an utterance or dialogue (such as a list, a question in an adjacency pair, deferential or uncertain utterance to a social superior);
    - falling tones *denote* a final element of a series of units in an utterance in a dialogue;
    - L* pitch accents *denote* focus operands in the initial part of a complex utterance;
    - H* pitch accents *denote* focus operands in the final part of a complex utterance;
    - so–called 'call contours' *denote* a dysfunctional channel (opening it, closing it or marking a communimcation breakdown).

It may be suggested that some, at least, of the beat (or baton), Butterworth and deictic gestures may be aligned with these prosodic categories. This is an empirical question and needs further investigation and more detailed semantic modelling.

One of the salient points of gesture, prosody and locutionary interrelations is *temporal organisation*. It is not only with respect to iconic and metaphorical gestures that similarities in temporal organisation constitute the basis for the semantic interpretation of gestural movements, but also the physical nature of gesture itself. The

semantic consequences of temporal relations between prosodic units and their temporal displacement with respect to their locutionary counterparts have been discussed by Kohler [12]. The temporal displacement of gestures with respect to their locutionary counterparts has also been discussed at several points in the literature; a detailed treatment with special attention to non–native communication, is given by Thies [21].

In [4] event logic and temporal calculi (e.g. the Allen relations) are used in order to construct a realistic model of phonetic and phonological representation which is similar to models used in the autosegmental gesture representations of Articulatory Phonology, and in order to create an operational model for processing these representations using Finite State Transducers.

Extensions of this modelling strategy to spatial transducers may be possible in order to relate gestures located at different reference areas of the body grid, as shown in Figure 4. Whether a more rigorous semantics (such as the fragments reviewed and developed in [16] or [11] for aspects of prosody) will become available for metalocutionary aspects of gesture remains an open challenge for future work.

References

[1] Jens Allwood. Cooperation and flexibility in multimodal communication. Gothenburg Papers in Linguistics, 1999.

[2] Jens Allwood. Bodily communication: Dimensions of expression and content. In Björn Granström, David House, and Inger Karlsson, editors, *Muldimodality in Language and Speech Systems*. Kluwer Academic Publishers, Dordrecht, 2002.

[3] Catherine P. Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49:155–180, 1992.

[4] Julie Carson-Berndsen. *Time Map Phonology*. Text, Speech and Language Technology. Kluwer Academic Publishers, Dordrecht, 1998.

[5] Dafydd Gibbon. *Perspectives of Intonation Analysis*. Lang, Bern, 1976.

[6] Dafydd Gibbon. Intonation in context. an essay on metalocutionary deixis. In Gisa Rauh, editor, *Essays on Deixis*. Narr, Tübingen, 1983.

[7] Dafydd Gibbon, Ulrike Gut, Benjamin Hell, Karin Looks, Alexandra Thies, and Thorsten Trippel. A computational model of arm gestures in conversation. In *Proceedings of Eurospeech 2003*, Geneva, 2003.

[8] Dafydd Gibbon, Inge Mertins, and Roger Moore, editors. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Kluwer Academic Publishers, New York etc., 2002.

[9] Dafydd Gibbon and Margret Selting. Intonation und die strukturierung eines diskurses. *LiLi, Zeitschrift für Literaturwissenschaft und Linguistik*, 49:53–73, 1983.

[10] Mark Andrew Keith Halliday. *Intonation and Grammar in British English*. Mouton, The Hague, 1967.

[11] Ekaterina Jasinskaja. Exhaustification and semantic relations in discourse. In Bart Geurts and Rob van der Sandt, editors, *Proceedings of the ESSLLI'2004 Workshop on Implicature and Conversational Meaning*, pages 14–19, Nancy, France, 2004.

[12] Klaus Kohler. Categorical pitch perception. In *Proceedings of the ICPhS, Tallin*, volume 11, pages 331–333, 1987.

[13] Mark Liberman and Ivan Sag. Prosodic form and discourse function. In *Proceedings of the Chicago Linguistics Society*, volume 10, pages 416–427, 1975.

[14] Craig Martell. Form: An extensible, kinematically-based gesture annotation scheme. In *LREC Proceedings*, pages 183–187, 2002.

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

[16] Arthur Merin and Christine Bartels. Decision–theoretic semantics for intonation. Technical report, Universität Stuttgart & Universität Tübingen, 1997.

[17] Cornelia Müller. The palm–up–open–hand: A case of a gesture family? In Cornelia Müller and Roland Posner, editors, *The semantics and pragmatics of everyday gestures*, pages 233–256. Weidler Verlag, Berlin, 2004.

[18] Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, and Janothers Henning. *HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An Introductory Guide*. Signum, Hamburg, 1989.

[19] John Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, 1969.

[20] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. Tobi: A standard for labeling english prosody. In *Proceedings of the 1992 ICSLPSociety*, pages 867–870, 1992.

[21] Alexandra Thies. First the hand, then the word: On gestural displacement in non-native english speech. Staatsexamensarbeit (M.A. equivalent), 2003.