Gesture Theory is Linguistics: On Modelling Multimodality as Prosody*

Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld,
Postfach 100131, 33739 Bielefeld, Germany
gibbon@uni-bielefeld.de

Abstract. Prosody and gesture have, with few exceptions, not appealed to computational linguists, and there is even less awareness of the parallels between them, though in modelling sign languages of the hearing-impaired, the linguistic metaphor of gesture patterns as 'phonology' has been very fruitful. On the other hand, many disciplines from anthropological linguistics to robotics are currently occupied very productively with conversational gesture studies, but nevertheless the field is still highly fragmented. Starting from (computational) linguistic models as metaphors for the forms and functions of gestural signs, it can be shown that similarities between gesture and speech (specifically: prosody), go much further than metaphors, opening up avenues for integrating speech and gesture theory and developing a shared ontology for speech and gesture resource creation and retrieval.

Keywords: gesture, prosody, speech technology, multimodality, resources.

1 Speech, gesture and technology

Gestures are an essential part of communication – not only the gesticulatory body language of everyday face-to-face communication and the signing of deaf communicators, but also in the production of speech and in the production of acts of writing, typing, manual morse code transmission, semaphoring and 'talking drums' and many other varieties of communication. In the broadest sense, music performance can also be seen as non-propositional gestural communication, though with such a generalisation about gestural communication one rapidly becomes overwhelmed with the dimensionality of the concept.

First, a note on gesture. Gestural communication, with few exceptions, remained the province of psychologists, sociologists, and of experts in the gestural sign languages of those with restricted hearing until relatively recently. In speech technology, the similarities between acoustic speech and sign language patterning, were recognised, and the technological development of multimodal systems for screen avatars and humanoid robots with speech and gesture communication developed. But computational linguistics has largely avoided the issue of gesture modelling, and until relatively recently gestures have not been the subject of computational linguistic investigations. This is one of the issues which will be taken up here.

Second, a note on prosody. The gestures involved in the production of prosody, the 'music' (rhythm and melody) of speech, are an intimate and essential part of verbal communication. The domain concerns the overall temporal coordination of all the gestures of speech, but it is the gestures of the larynx (which control the vocal cords) which are the key contributors to the prosodic domain. Linguistics (particularly applied linguistics, both in the domain of foreign language teaching and clinical linguistics) has been concerned with prosody for many decades,

* Supported in part by DAAD grant to *ModeLex* project 2001-2004 (FG 'Task-oriented communication'). Discussion of parts of this work with GESPIN conference participants is greatly appreciated.

and there are many standard works dealing with prosodic features of speech, as well as innumerable journal and conference articles on different aspects. Speech technologists have become increasingly interested in prosody over the past two decades, partly from the traditional linguistic perspective that the 'music of speech' is an important indicator of the structure of spoken utterances and of their status as speech acts and dialogue acts, partly simply from the point of view of the naturalness and acceptability of artificial speech. But another influence during the past five years or so is the realisation that communication is not only rational and representational (the fields of traditional syntax and semantics), but also intuitive, emotional, expressive and appellative, functions which have traditionally been regarded as non-linguistic, or at most of psycholinguistic interest. Cutting edge research in this area has been done by Cambell (2007), who has looked beyond the classic 'rhythm and melody' concept to the paralinguistic vocalisations, which he calls 'grunts', and by Fischer (2000) on discourse particles, and Tseng (1999), who have looked at the dysfluent structures in non-fluent communication and their properties. Gibbon (2009) has drafted a first ontology for the description of spoken language, which is in large measure also applicable to gesture.

The present contribution addresses some of the transitional work required in order to regard the fields of gestural and prosodic communication in an integrative perspective. First an outline of the functions of gesture is given, then the formal properties of functions of gesture are discussed, based on a priori prosodic models.

2 Gestures

2.1 Examples

We can initially delimit the meaning of 'gesture' in a narrower and tractable sense, by means of a few examples. Churchill's use of the victory sign, the Roman victory salute, an air kiss to a parting close friend, a wave, beckoning with a finger, a dismissive hand movement, cupping a hand around the ear, the f-sign or 'the finger': these gestures are, at first glance, simple to categorise: they are all hand movements with some communicative function; they are all signs. But they are different, and the details of the communicative functions also need to be categorised, as well as the similarities, and this is perhaps not as simple as it may seem.

The heuristic for approaching gesture in the present context is to use linguistic models rather than psychological or sociological ones. That is, gestural communication is treated formally and empirically pretty much in the same way as spoken language, with the exception of a switch from auditory to the visual modality. The forms, structures and functions of gestures will be taken to be analogous to the forms, structures and functions of speech. Specifically, the relation of gestures to speech will be taken to be analogous to the relation of prosody as temporally parallel to speech. On this basis, a systematisation of gesture from paradigmatic (classificatory, taxonomic) and syntagmatic (compositional, mereonomic) perspectives will be proposed, though not all details of the speech-gesture parallels can be dealt with in the present context.

From the perspective of the linguistics of discourse, all of the gestures mentioned at the start either initiate or terminate a phase of phatic interaction, either starting or ending some kind of communicative encounter or sub-encounter (i.e. episode in the encounter). As a first approximation to characterising the differences in the forms and functions of gestures, a standard strategy of *definitio per genera proxima et differentia specifica*, can be followed, and could run something like this (here concentrating only on manual gestures):

> A communicative gesture is a movement of the hands,
> (a) with one or both hands, hand/arm/fingers in a certain shape *A*, with (no) contact with another part of the body, with no contact with an interlocutor, starting at point *B*, finishing at point *C*, and
> (b) with positive or negative associations or sanctions, initiating/terminating a dialogue act *D* with a goal *E*, in a social configuration *F*.

The differentiating properties listed in the definition constitute dimensions in a quality space which can be represented formally as an attribute value structure. This definition is at a relatively coarse level of granularity, and much more detailed specifications are required.

2.2 Multimodality

Face-to-face communication takes place in several different modalities and sub-modalities, illustrated in stylised fashion in Figure 1. The Figure shows five modalities which can be used to form parallel communication channels: facial-visual, oral-visual, oral-auditory, hand-visual, foot-auditory. Hand-auditory and foot-visual could have been added.

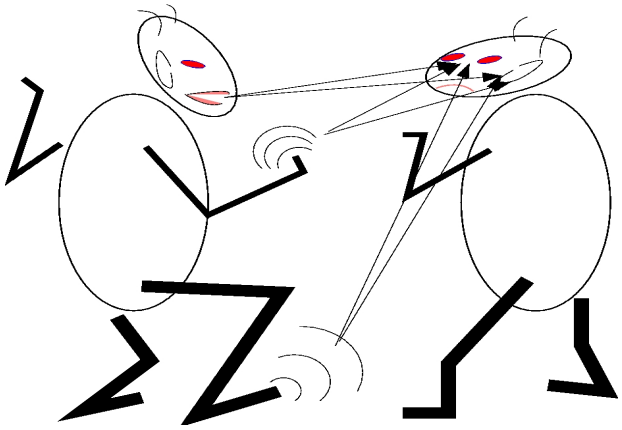

Figure 1: A stylised view of parallel multimodal channels.

The concept of 'modality' needs to be distinguished from that of 'medium', and the concept of 'sub-modality' needs to be introduced, as well as the notions of 'multimodal system' and 'multimedia system' which are used in speech technology (see also Gibbon 2000:105).

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

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

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. It is irrelevant for this definition whether technical channels intervene. The voice is one modality, the limbs and other body areas and movements represent other modalities.

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. Example: 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 in one of the visual modalities are text and still or animated images.

2.3 A basic semiotic model

It is not sufficient to characterise communicative gestures by means of their forms alone, since they have a semiotic function which interprets an atomic or structured gesture (which can be

thought of as a cognitive unit) in terms of both a meaning and a form (which can be interpreted as two domains of reality, the latter being a proper subdomain of the former (Gibbon 2006):

$\text{Interpretation}_{\text{semantics}}$:	gesture →	cognitive or physical meaning
$\text{Interpretation}_{\text{modality}}$:	gesture →	sensory form

For example, the function of a waving gesture is to attract attention to the beginning or the end of a social contact, and the form of a waving gesture is a certain movement of hand and arm, perhaps accompanied by eye contact and specific head and body movements.

Thus:

$\text{interpretation}_{\text{semantics}}$:	wave	→	attention attraction at encounter periphery
$\text{interpretation}_{\text{modality}}$:	wave	→	iterated left-right / up-down hand/finger motion

The following sections will deal with the functions of gesture, then the forms of gesture and the forms of prosody.

3 Gesture functions

In linguistics, a number of systematic models for the functions of speech and text have been proposed, which can equally well be applied to gesture. The most well-known of these is the old and simple model of Bühler (1934): communication takes place in a quadruple of objects with three (possibly overlapping) subsets known as 'constitutive factors': *Speaker*, *Addressee*, *Context*, *Sign*, and three functionalities of signs are defined as functions from the set of signs into these sets:[1]

expressive function :	*Sign*	→	*Speaker*
representational function :	*Sign*	→	*Context*
appellative function :	*Sign*	→	*Addressee*

This simple model has been extended by many scholars. Jakboson (1960) used slightly different terminology for his extension:

expressive function :	*Message*	→	*Sender*
representational function :	*Message*	→	*Context*
conative function :	*Message*	→	*Receiver*
phatic function:	*Message*	→	*Contact*
metalingual function:	*Message*	→	*Code*
poetic function:	*Message*	→	*Message*

More recently, Allwood (2002) used the factors *Sender*, *Recipient*, *Expressions*, ('sign'), *Media*, *Content*, *Purpose*, *Environment*, though factors such as 'purpose', 'content', 'environment', 'sender' clearly have an entirely different ontological status from those of the Bühler and Jakobson models; these are ontologically more homogeneous. Allwood subdivides the representational function (the semantic domain) according to the classic Peircean semiotic model into 'symbol', 'icon' and 'index' functions. In the present study, other specifically semantic functions such as naming, predication, quantification and metaphor are introduced as functions of gesture.

In the pragmatic domain (the expressive and conative functions) a similar strategy is applied here, with reference to well-known pragmatic functions of speech, such as dialogue management, speech-acts, turntaking, backchannelling.

A well-known related approach, which is, however, restricted to semantic and expressive functions of gesture, is given by McNeill (1992). The following list adds two further function categories (*emblems* and *affectives*) to McNeill's five main categories:

[1] I am formulating this more formally than Bühler does, and interpreting his relations as functions.

1. *Iconics*, where the gesture resembles the referent (e.g. describing an action or shape of an object with the hands).
2. *Metaphorics*, where the *vehicle* (the gesture) relates in one of a number of metaphorical ways to the *tenor* (non-literal meaning) of the gesture, e.g. indicating a container or conduit for ideas, or a gift of an idea or suggestion (cf. Lakoff & Johnson 1980).
3. *Beats*, where the hand, head, eybrows move roughly in synchrony with the rhythm of often emphatic speech, mark a sequence, or a hiatus such as a change of theme or focus.
4. *Cohesives*, which create a gestalt in gesture space which is coextensive with a spoken utterance or – hierarchically – with its parts.
5. *Deictics*, which may indicate an actual physical position, size, distance or direction, but may also place concepts metaphorically in physical gesture space
6. *Emblems*, which are fairly highly conventionalised, lexicalised gestures, and constitute the most well-known type of gesture.
7. *Affectives*, which display emotional states and events.

The phatic greeting and farewell gestures noted at the beginning of this contribution belong to the category of emblems. Functionally, these phatic gestures are like interjections, in that they have a relatively fixed but often hard to define form-function relationship, and they do not fit into the regular flow of speech, but have an autonomous attention-getting, channel-creating or emotional status. The same applies to the chant-like stylised phatic intonation (Gibbon 1976) used in calling, routine lists and corrections, and with some interjection-like greetings ("Hello-o!") and farewells ("By-ye!"). Other gesture emblems are more clearly related to the main parts of speech of a language and, like other parts of speech, are highly language-specific or culture-specific: various configurations of hand and fingers with a wide range of clearly identificable meanings such as success, pleasure, idiocy, cuckoldry, disgust, eating, drinking and telephoning.

Note that McNeill's categories are not necessarily mutually exclusive types, but rather parameters whose values can co-occur in any given gesture: emblems can be iconic and metaphorical, for example, holding the hands wide apart to indicate the great importance of some issue.

In terms of the previous categorisations, the functionality of the iconics, metaphorics, deictics and emblems is semantic, i.e. representational; the functionality of affectives is pragmatic, i.e. expressive; the functionality of beats and cohesives is structural or syntactic. The functionalities of gestures are, however, much broader, and in order to characterise these, more highly differentiated linguistic and computational linguistic approaches are required. Current discussion is based on 'dialogue acts', which are domain-specific versions of speech acts.

In speech act theory 'illocutions' such as stating, questioning, commanding, but also task-specific institutional acts such as baptising, marrying, judging, are described in terms of the contextual conditions which a speaker needs to fulfil in order to communicate successfully. It is not possible or necessary to go into details here, but Searle's classic set of nine conditions (1969) for promising can be cited, here in informal formulations. Normal input and output must obtain, and the following kinds of information must be expressed: a proposition; a future act of the speaker; a preference of the hearer for the speaker to do this act (as opposed to a warning!) and the speaker believes this preference; that the act might be done anyway is not obvious; the speaker intends to do the act; the speaker intends to have an obligation to do the act; the speaker intends the hearer to realise that by uttering the promise the hearer recognises that the act would not be done anyway and that the speaker undertakes an obligation; the semantic rules of the shared language define a correct and sincere utterance of a promise on the basis of the preceding eight conditions.

In his *Dynamic Interpretation Theory* of discourse, Bunt (2000) extends the speech act approach into a broad and very detailed functional categorisation which is intended for the practical annotation of dialogues for computational corpus linguistic analysis. So far, this kind

of detailed annotation has not been applied to the analysis of gestures, and it is not possible to give all the details here; the main categories will be sufficient to illustrate the point.[2] The key distinction in Bunt's approach is between generic ('general purpose') dialogue acts, which essentially cover the speech acts of traditional speech act theories such as that of Searle (1969), which are applicable to all communicative situations, and situation-specific ('dimension specific') dialogue acts, which are characteristic of specific kinds of interaction. The distinction is interesting not only as a basis for a systematic dialogue act ontology, but also because of a presumption it appears to make between universal and culture-specific speech acts, 'culture-specific' rather than 'language-specific' because dialogue acts are concerned with the use, not the form of language.

Table 1: Application of Bunt's main Dynamic Interpretation Theory (DIT) categories to gesture description.

Dialogue act functions	Gestures
1. General Purpose communicative functions	
1. Information transfer	
1. Information seeking	Querying gestures (e.g.raised eyebrows)
2. Information providing	Raised or wagging finger ('didactic finger')
2. Action discussion functions	
1. Commissives	Promise, contract (e.g. handshake)
2. Directives	Dismissal (e.g. sideways hand wave)
2. Dimension-specific communication functions	
1. Activity-specific functions	
1. Open meeting	e.g. beat table with gavel
2. Bet	e.g. handshake
3. Congratulation	e.g. handshake, pat on shoulder/back
4. ...	
2. Dialogue control functions	
1. Feedback	e.g.nod, head shake
1. Auto-feedback	'Thinking gestures', e.g. finger mouth
2. Allo-feedback	e.g. nod, head shake
2. Interaction management	
1. Turn management	e.g. raise/fall of hands, eye gaze
2. Time management	e.g. beat gestures
3. Contact management	e.g. wave hands
4. Own communication management	Error flagging, e.g. sideways hand wave
5. Partner communication management	Attentiveness, e.g. raised eyebrows
6. Discourse structure management	Topic shift, e.g. hand gestures
7. Social obligations management	
1. Salutation	e.g. wave, salute, air kiss, cheek kiss
2. Self-introduction	e.g. bow, handshake
3. Apologising	e.g. prayer gesture
4. Gratitude expressions	e.g. thumbs up gesture
5. Valediction	e.g. wave, handshake

4 Gesture forms

The starting point for systematising the parts and combinations of gesture forms is the insight that a single gesture has an identifiable sequence of phases (Kendon 1996), as already noted:

> A gesture is a clearly demarcated symmetrical movement from a rest position via a peak (centre or stroke) back to a rest position.

Gestures defined in this way are primarily *atomic gestures*, which are segments in the *stream of gestures* in the same sense that morphs in the *stream of speech* are segments. The

[2] For definitions see: `http://let.uvt.nl/general/people/bunt/docs/dit-schema3-2.html`

structure of an atomic gesture is analogous to that of the sonority curve of a prototypical CVC syllable in speech: from a well-demarcated low sonority initial consonant through a high sonority vocalic segment to a low sonority final consonant. Since gestures, unlike syllables per se, have meanings, they are more analogous to monosyllabic morphs (realisations of morphemes) than to syllables, and a fortiori, at an appropriate level of abstraction, to inventarisable morphemes. Atomic gestures may thus be said to form a basic vocabulary which is formally comparable to the vocabulary of morphemes of spoken language.

But the issue of the syntax of complex gestures arises, with two compositional issues: Is there a grammar of gesture sequences? Is there a grammar of gesture synchronisation? The second issue is formally the same as that of assigning prosody to locutions as a parallel channel with semi-independent forms and functions, corresponding to the basic heuristic of the present approach. Based on the present strategy of using linguistic analogies, a number of gesture compositionality issues require treatment:

1. *Composition of atomic gestures* as the combinatorics of the simultaneously occurring features which represent the paradigmatic relations between atomic gestures. Some examples of these combinations were given in characterising paradigmatic relations between gestures.
2. *Sequential combinatorics* of atomic gestures, including whether gesture sequences are only 'flat' or linear or whether there are perhaps also hierarchically structured sequences; on very general kind of sequence moves from a wave to start a conversation, via the conversational gestures of the actual interlocution, to a wave at the end. That the sign languages of the hearing-impaired have intricate syntax, like the syntax of speech, is well-known.
3. *Synchronous combinatorics* of atomic gestures, both with each other (e.g. waving and smiling at the same time) and with speech (e.g. pointing gestures together with deictic expressions, emphatic gestures together with emphasised words, overall cohesive gesture gestalts coextensive with utterances).
4. *Word vs. sentence combinatorics*, i.e. whether there is a principled distinction between gestural 'sentences' and gestural 'compound words'. For instance, some gestures consist of two distinct movements in sequence, like the Roman Catholic gesture of crossing oneself, with an assimilation effect of not returning to a rest position between the vertical and the horizontal gesture parts. Gestures of this kind could be described as 'gestural words' or alternatively as 'bi-atomic gestures'. This gesture type contrasts with a combination of two independent deictic gestures, e.g. the index finger pointing to a person, then (or simultaneously, using two hands) the thumb pointing to the door, meaning "You, get out!"

Two complementary recent formal studies of gesture syntax are available:

1. CoGesT: A study of the basic combinatorics of type 1 is given in Gibbon & al. (2003), where a formal grammar for hand gestures is provided. A basic distinction is made into Simplex Gestures and Compound Gestures. Simplex gestures are of two types: *2-place static* (where a gesture with a hand configuration is held) and *9-place dynamic* (with specification of Source (Location and Handshape), Trajectory (Lateral, Sagittal and Vertical Direction; Shape, Form, Size and Speed), and Target (Location and Handshape). These attributes are represented as a vector, which may be enhanced with specifications for two-member gestures (e.g. symmetric, where hands make mirror image movements, or parallel, where hands make the same movement) and indicators for the left or right side of the body for paired members, e.g. left or right hand. The model also has a specification for iterative gestures such as waving.
2. MURML: A set of XML conventions (Wachsmuth & Kopp 2002) for representing gestures in a robotics context, with specifications for Timeline, Symmetry, HandShape, PalmOrientation, ExtendedFingerOrientation, HandLocation, ShoulderLocation,

CentreLocation, Start, Direction, Distance. The MURML specification has very many more details for specification of attributes of the hand than can be discussed here.

The basic Source-Stroke-Target structure is represented differently in the two approaches: in CoGesT, the structure is represented directly by the Source-Trajectory-Target triple, while MURML uses a Source-Direction-Distance format which is more convenient for calculations in the robotics environment in which it is located, but does not give the shape of the trajectory.

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
    1. the direction of a movement, which is given in a vector for all three axes relative to the previous location,
    2. 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,
    3. the shape of the hand during the movement,
    4. a description of the size of a gesture and the speed of the movement,
    5. 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. The CoGesT vectors could easily be described by a regular grammar or finite state automaton: because 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 grammar does not explicitly represent the distinction between sequential compositionality and simultaneous compositionality, which will be discussed below. This distinction must be specified for each rule. From the point of view of generality, this is not completely satisfactory; there remains much to be done.

The third compositionality issue, synchronisation with the utterance, is handled by the timeline in the MURML notation, but requires more detailed temporal models for a full description. A suitable basis for an explication of gesture synchronisation, whether for speech or non-speech gestures, is the notion of *Time Type* (Gibbon 2006), representing levels of abstraction from physically measurable time:

1. *Categorial Time*: time is specified simply as an abstract property or category, such as *duration* or concatenation representing sequences in linguistic descriptions.
2. *Relational time* (*rubber time*): time is specified as precedence and overlap relations as used in Autosegmental Phonology, and formalised in van Benthem's Event Logic and in Allen's Interval Calculus (cf. Carson-Berndsen 1998).
3. *Absolute time* (*clock time*): time is specified as a set of measuring points, as in recordings and annotations of digitised audio and video signals.

The variety of temporal relations between gestures, prosody and speech may be represented using Allen's Interval Calculus. The Interval Calculus defines the thirteen possible relations between two intervals X and Y (see Figure 2). Using this approach, a gesture can be defined as a pair of a *movement* and an *interval*, $G = < M, I >$, and gesture synchronisation can then be defined as a relation between the intervals X and Y in a set of such gestures: $SYNC(G_X,G_Y)$.

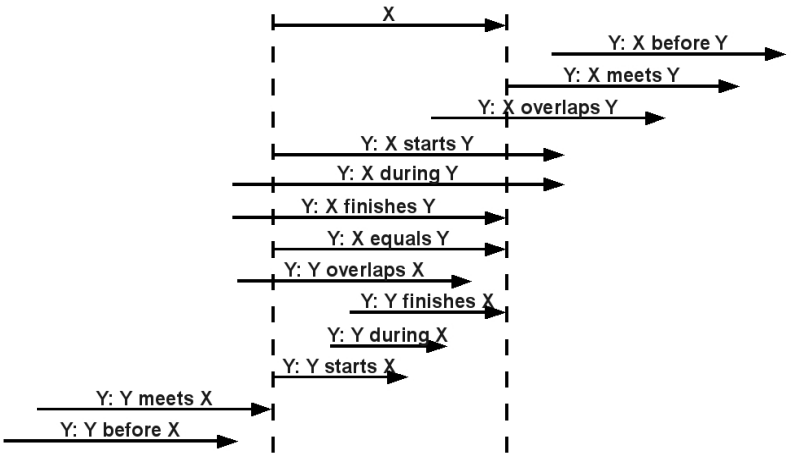

Figure 2: Illustration of interval relations in Allen's Interval Calculus.

Carson-Berndsen (1998) has shown how interval and event structures of the kind shown in Figure 2 can be formalised within the Time Type framework as finite state transducers which map between Time Types. Thies (2003) has shown empirically that for certain types of gesture there is a displacement relation: the synchronisation relation between a hand gesture and an associated word constituent is typically, in terms of Allen interval relations, either $OVERLAPS(G_{HAND},G_{WORD})$, $BEFORE(G_{HAND},G_{WORD})$ or $MEETS(G_{HAND},G_{WORD})$.

5 Conclusion

In the present discussion, the issue of integrating communicative gesture with locutionary levels of speech was addressed, and for this purpose, functional and formal categories which are normally used for linguistic analyses of these locutionary aspects were used, rather than the more ad hoc categories used in the diverse literature on gesture analysis. The main clue to integrating gestural and locutionary communication lies in a similarity to prosody.

On the functional side, it turns out that to describe the pragmatics of gestural communication quite sophisticated categorisations of dialogue acts are required; previous descriptions of the functions of gesture were rather simple in comparison, and restricted to a few semantic and structural categories.

On the structural side, unlike the core structures which are traditionally analysed in linguistics and modelled in computational linguistics with concatenative calculi, prosody requires the incorporation of concepts of 'overlap', 'simultaneity', or of 'parallelism' in a domain-oriented sense, not just in the sense of breadth-first parallel search.

On the basis of the linguistic analogues discussed in the present contribution, the issue of modelling gesture in relation to the locutions and prosody of speech would seem to be reasonably clear. In order to achieve a descriptively adequate, and technologically applicable model, however, these descriptive categories need to be applied to the annotation of large corpora of gestural data – an expensive, time-consuming and challenging, but worthwhile task.

References

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

Bühler, Karl. 1934. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Jena: Gustav Fischer.

Bunt, Harry. 2000. Dialogue pragmatics and context specification. In H. Bunt and W. Black, eds., *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. John Benjamins, Amsterdam, pp. 81–150.

Campbell, Nick. 2007. On the Use of NonVerbal Speech Sounds in Human Communication. *COST 2102 Workshop (Vietri) 2007*, pp. 117-128.

Carson-Berndsen, Julie. 1998. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. New York: Kluwer Academic Publishers.

Fischer, Kerstin. 2000. *From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Particles*. Berlin: Mouton de Gruyter.

Gibbon, Dafydd. 1996. *Perspectives of Intonation Analysis*. Bern: Lang.

Gibbon, Dafydd, Inge Mertins and Roger Moore. 2000. *Handbook of Multimodal and Spoken Dialogue Systems*. Dordrecht: Kluwer Academic Publishers.

Gibbon, Dafydd. 2005. Prerequisites for a Multimodal Semantics of Gesture and Prosody. In Harry Bunt, ed., *Proceedings of the International Workshop on Computational Semantics* 6.

Gibbon, Dafydd. 2006. Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data. In Stefan Sudhoff et al., *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter, pp. 281-209.

Gibbon, Dafydd. 2009. Can there be standards for Spontaneous Speech? Towards an Ontology for Speech Resource Exploitation. In Shu-Chuan Tseng, ed., *Linguistic Patterns in Spontaneous Speech*. Language and Linguistics Monograph Series A25. Taipei, Taiwan: Institute of Linguistics, Academia Sinica.

Gibbon, Dafydd, Ulrike Gut, Benjamin Hell, Karin Looks, Alexandra Thies and Thorsten Trippel. 2003. A computational model of arm gestures in conversation. *Proceedings of Eurospeech 2003*, pp. 813-816.

Jakobson, Roman. 1960. Linguistics and Poetics: Closing Statement. In: Thomas Sebeok, ed., *Style in Language*

Kendon, Adam. 1996. An agenda for gesture studies. *The Semiotic Review of Books*, 7.3:7-12.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

McNeill, David. 1992. *Hand and Mind*. Chicago: University of Chicago Press.

Searle, John. 1969. *Speech Acts*. Cambridge: Cambridge University Press.

Tseng, Shu-Chuan. 1999. *Grammar, prosody and speech disfluencies in spoken dialogues*. Dissertation, Universität Bielefeld.

Thies, Alexandra. 2006. *First the Hand, then the Word: On Gestural Displacement in Non-Native English Speech*. MA equiv. thesis, Universität Bielefeld.

Wachsmuth, Ipke and Stefan Kopp. 2002. Lifelike gesture synthesis and timing for conversational agents. In Ipke Wachsmuth and Timo Sowa, eds., *GW 2001, LNAI 2298*:120-133. Berlin: Springer Verlag.