Poznań Studies in Contemporary Linguistics 47(3), 2011, pp. 470–508

doi:10.2478/psicl-2011-0026

DE G

VERSITA

MODELLING GESTURE AS SPEECH: A LINGUISTIC APPROACH

DAFYDD GIBBON
Universität Bielefeld
gibbon@uni-bielefeld.de

ABSTRACT

Gesture communication, like prosody and paralinguistic voice features, strikes the attention when there is too little of it, too much of it, or when it does not seem to fit the words or the situation. The present study follows the principle that gesture is similar to some aspects of speech, particularly prosody and parts of the lexicon. Description of visual gesture articulation is therefore treated as a conservative extension of descriptions of vocal speech gesture articulation. Well-tried models of speech forms and functions are deployed, together with accounts from gesture studies from psychology to robotics. Evidence is taken from video data of story-telling in Ega, an African language, and in German, and the adequacy of descriptive and computational models of the forms and functions of speech is discussed, with a proposal for the formal modelling of speech-like timing of gesture articulators by means of *Time Types* in the *Linear-Feature-Timing-Realtime* (*LFTR*) model. Finally, an integrative model for combining visual and vocal gesture articulations into a comprehensive functional model of multimodal communication is proposed: the *Rank Interpretation Model* (*RIM*).

KEYWORDS: Gesture; prosody; speech; linguistics; computation.

1. Integrating visual and vocal gesture[1]

Conversational gesture modelling in the present study starts from linguistics rather than from the disciplines which have traditionally been most concerned

[1] This contribution is a considerably modified and extended version of the paper presented at the GESPIN conference in Poznań, September 2009. I am grateful to participants in this conference for comments, particularly Adam Kendon and Nicla Rossini. The research was partly funded by the projects "EAGLES II" (European Commission), "Modelex: Theorie und Design Multimodaler Lexica" (Deutsche Forschungsgemeinschaft), "Ega: Documentation of an Endangered Ivorian Language" (VW Foundation).

with analysis and modelling of the gesture domain: psycholinguistics, sociolinguistics, body language training and, more recently, areas of applied informatics such as robotics and video game development. The present research has been partly informed by these disciplines, but also by a long series of studies on prosody and the lexicon which have gradually suggested with increasing insistence that gesture and speech are two of a kind, and that in particular conversational gestures are related more closely in form and function to prosody as well as to parts of the lexicon than has previously been thought.

The central thesis of the present study is that "gesture is a linguistic domain". Gesture is obviously also within the purview of many other disciplines, but the linguistic dimension has been rather neglected. The aim of the present research is thus to connect linguistics and gesture studies in a new way, by conservatively extending a suitable set of descriptive and computational linguistic models to cover gesture. One interest of this "linguistic turn" is that the "gesture is a linguistic domain" principle represents a null hypothesis which brings with it a clear burden of falsification or confirmation. But then for the past century and a half speech has been modelled as gesture in articulatory phonetics, and even written texts are stored traces of visual gesture, both in form, producing character configurations, and functionally: "he signed the document with a flourish". So, at least on circumstantial evidence, the mining of descriptive and computational linguistic approaches for both visual and vocal gesture description appears *prima facie* to have a reasonable chance of success.[2]

Linguistically informed approaches to gesture study are not exactly new (cf. Pittenger et al. 1960; Birdwhistell 1970; Kendon 1972 *et passim*; McNeill et al. 2001; Gibbon et al. 2003; Trippel et al. 2004; Gibbon 2005; Rossini 2004, in press). There are precedents elsewhere, in the "phonology" metaphor for languages of the hearing-impaired (Brentari 1998), in a century and a half of modelling speech gestures in phonetics and more recently in Articulatory Phonology (Browman and Goldstein 1992). To emphasise the relationship, the body members involved in all gestures, vocal and visual, will be referred to as "articulators", as in articulatory phonetics, and their movements will be called "articulations".

The present approach partly contrasts with and partly overlaps that of McNeill (1992, 2005), in which gesture is fundamentally not morphemic, not

[2] Many thanks to an anonymous reviewer who (perhaps jocularly) adamantly misrepresented the model-based deductive reasoning of the present study as an ignorant "attempt to authoritatively legislate" about gesture studies (while condoning the useful taxonomies in the study), and who thereby triggered much useful additional clarification.

compositional, and not defined in terms of form–meaning conventions, but rather holistic or global, i.e. the meanings of parts are determined by meanings of the whole within the overall context. The holistic premise also works for cases of context-dependent meaning, idiomaticity, ambiguity and vagueness in speech. However, the core of the linguistic approach is fundamentally analytic: the meaning and form of the whole are functions of meanings (and forms) of parts. In this respect, the linguistic approach relates more closely to the work of Kendon (1992, 1996, 2004).

Speech is vocal gesture transduced into sound. Two linguistic components are focussed as productive models for both vocal and visual conversational gesture: the *prosody* component (intonation, accent and rhythm) on the one hand, and the *lexicon* component on the other. In the domain of visual articulations (some of which are also transduced into sound), a non-absolute distinction must be made between conversational gesture or gesticulation on the one hand, and the signing or sign languages of the hearing-impaired on the other. Signing relates closely to the locutionary component of speech, while conversational gestures relate mainly to prosody. Lexical gestures for relatively limited sets of culture-specific actions and objects are a special case. For "prosodic gesture", the following proportionality of "prosody is to gesture as locutions are to signing" is maintained as a working hypothesis:

prosody : gesture :: locution : signing

The "prosody : gesture" relation has long been a *topos* in the field, figuring in pioneering work by Birdwhistell (1970) and Kendon (1972). A corollary of proportionality is that speech and conversational gesture are as like as chalk and cheese: conversational gesture is a *comparandum* for prosody, not for speech as a whole. Prosody is the relevant linguistic discipline (cf. Gibbon 1976a; Gibbon et al. 1984). It will be shown in the following discussion that gestures in general share the functionality and gradient characteristics of prosody, of paralinguistic features and of lexical interjections, rather than of speech in general.

It is clearly a triviality to maintain that "all communicative acts are gestures". Nevertheless, the generalisation is a useful reminder of commonalities. Clapping, stamping, snapping fingers, whistling, vocal articulations in speech are acoustically transduced gestures. Silent signalling in secretive situations and semaphoring in acoustically hostile environments involve gestures. Morse code, handwriting and typewriting/keyboarding use highly structured manual gestures. Indeed, the general communicative significance of the concept of gesture

is generalised in the metaphorical use of "gesture" to mean a kind (though "symbolic") act.

Most available gesture studies, including the present research, focus on the analysis of gesture productions (cf. Kendon 2004; McNeill 2005, and references there). However, perception of the causal effects of gestures in different modalities would be equally deserving of empirical study: the act of communication proceeds solely via the production of a sign as a gesture form, the transmission of the form through some medium and the perception of the sign through the form in context, and not through hermeneutic magic.

The following section clarifies basic concepts of modality, media and multimodality, with an example of the "phonetics" of beat gestures and an initial characterisation of "gesture". The subsequent sections delimit and characterise the domain of visual gesture, models of the functions of gesture, and explicitly formal models of the forms of gesture. Finally conclusions about principles of modelling vocal and visual gestures in integrated fashion are drawn.

2. Gesture, modalities and media

The topic of this section is the relation between visual, vocal and other modalities, and between the methods for analysing them. Parts of the methodological discussion are an excursus on "phonetic" evidence and measures for a type of visual gesture ("beat" gestures), and a discussion of the semiotic status of gesture.

2.1. Modalities, media and the "phonetics" of gesture

The first keys to distinguishing communication modes such as conversational gesture, signing, semaphoring, writing and speaking lie in properties of the medium. Movements of articulators in signing and gesticulation are directly perceived in the visual medium. Semaphoring is gesture in which the movements of articulators are amplified technically by the use of flags. Hand-shaking involves articulators which touch the addressee. Writing is gesture in which the articulators leave visible traces on paper, stone, wood, screens etc. generally by means of a tool which may be simpler (e.g. pen and ink, pencil, sharp point) or more complex (e.g. computer), or by means of dictation to another person using such tools. Vocal gestures (and similarly clapping, snipping, stamping, whistling), are movements of articulators which generate audible sounds which are

used in single modalities or in multimodal and multimedial configurations (cf. Figure 1).

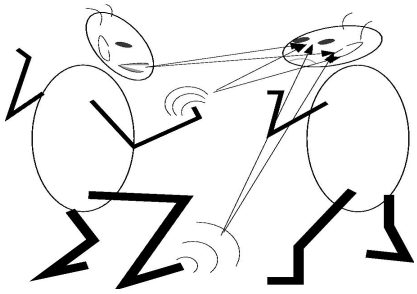

Figure 1. Caricature of parallel vocal and visual modalities and submodalities.

The terms "multimodal" and "multimedia" are often confused. Multimodal communication occurs via more than one modality (e.g. oral–auditory, manual–visual), while multimedia communication takes place via different technical media, in subchannels which may be inserted simultaneously and sequentially into the human communication channel (Gibbon et al. 2000). The following definitions clarify the modality-medium distinction:

> A *modality* is a communication channel characterised by a pair of human motor output and sensory input organs.
>
> A *medium* is a face-to-face or technical channel within a modality.
>
> A *multimodal complex* is a combination of modalities (such as the combination of visual and vocal gestures).
>
> A *submodality* is a use of a modality which is functionally distinct from other uses, e.g. prosody in speech, or hand gestures in visual gesture complexes.

Multimedia communication may either be in the same modality (e.g. speech and music; text, photographs and video) or may transduce from one output modality

(e.g. visual text) into another input modality (e.g. acoustic speech). The form of a communicative act is accordingly modelled by a causal chain in which a *gesture* is transformed within the constraints of the *medium* into a *signal* and then a *percept*:

$$ARTICULATION(articulator_i) \rightarrow ... SIGNAL(medium_j) \ldots \rightarrow PERCEPT(sense_k)$$

A more detailed version of the causal chain model is familiar from phonetics, with its three main domains of *articulation* (the gestures)*, transmission* (in acoustic media), and *perception* (pattern recognition). The devil is in the detail, of course: the *ARTICULATION* function may be somewhat similar for vocal and visual gesture, but the *SIGNAL* and *PERCEPT* functions differ greatly in the different modalities, and may or may not thereby condition differences in semiotic function. The integrative linguistic approach argues that they do not, but that the modalities are specialisations of the same communicative process, constrained in relatively superficial modality-specific ways.

2.2. Evidence: A "phonetic" analysis of beat gestures

Evidence for gesture data is characteristically provided by video films and photos (epistemically comparable with phonetic audio recordings, video recordings of articulations and graphic signal transformations). Like transcriptions and analytic verbal descriptions, the iconic line-drawings often found in the literature (e.g. Kendon 2004) are not directly evidential as they involve massive interpretation by much greater selection, abstraction, and stylisation than recordings. Functional descriptions of gestures are necessarily interpretations.

The scenario to be described in "phonetic terms" involves direct evidence for beat gestures in traditional story-telling in Ega (Niger-Congo, Kwa, Côte d'Ivoire, ISO 639-3: *ega*). Figure 2 is a frame set from the peak excursion of right-hand beat gesture with large upward excursion (cf. also Rossini and Gibbon 2011). The narration is punctuated by sequences of numerous iterated beat gestures at a "heartbeat" rate averaging around 78 per minute, comparable with accents (stresses) in speech. Beats have, intuitively, the function of manifesting rhythmic coherence, like prosodic accentuation and rhythm. Accent–beat synchronisation has been sporadically investigated from Lashley (1951) to Cummins and Port (2006) and Rossini (in press). A frequently used phonetic measure for regular temporal patterning was therefore chosen, the *normalised Pairwise Variability Index*, *nPVI*, a useful but not uncontroversial measure (cf. Gibbon

2006 for technical discussion). The *nPVI* for a given utterance is a function of the averaged differences between interval durations of adjacent segments (phonemes, syllables, feet, etc.) and ranges from 0 to an asymptote of 200. In the present case, the adjacent segments are inter-beat intervals.

Figure 2. Still frame sequence with large right-hand baton gesture (240 ms, Ega conteur Grogba Marc).

The recording was annotated for beat gestures of the hands and arms using the *Anvil* video annotator. For present purposes, left-hand, right-hand and synchronised beats were combined. Time-stamps and duration measures were extracted, and the *nPVI* applied:

min = 80 ms	*max* = 3200 ms	*range* = 3120 ms
mean = 770 ms	*sd* = 730 ms	*nPVI* = 5

Bearing in mind that the *nPVI* for speech units is generally between about 30 and 70, the very low value of 5 indicates extreme mean regularity, confirming intuition about beat rhythm. However, the range and standard deviation are actually rather high, which seems to belie the straight *nPVI* value and indcate regularities of different temporal scope. This demands explanation, for which the beat amplitude-timing relation will be briefly described.

Visualisation of the timeline with annotations of beat excursions on a scale from 1 (small) to 4 (very large) confirms that the highly regular *nPVI*, taken together with the large ranges and standard deviation, generalises over temporal clustering of different domains: there are different, possibly hierarchical strata of beat rhythm patterns (Figure 3; the thickness of the lines in the figure has no significance). The closer-spaced beats tend to sync with syllables (Ega tends to be syllable-timed), more broadly spaced beats tend to sync with emphasised

words. Slopes of lines between peaks of given heights are a function of beat amplitude differences and speed of repetition. Flatter slopes after higher peaks tend to indicate pauses and the end of major discourse units. A full analysis of the hierarchical gesture rhythm structure of the narration on the lines of existing phonetic analyses of speech rhythm is beyond the scope of the present discussion, but the present analysis nicely illustrates the utility of phonetic methods in gesture studies.

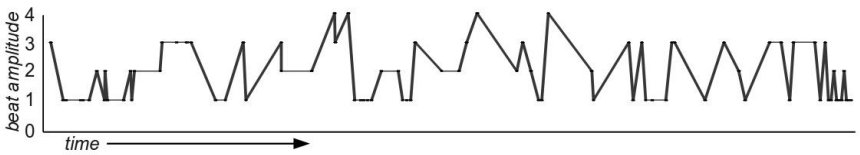

Figure 3. Beat gesture timeline in the first 60 s of the Ega story, showing irregularities and subregularities.

2.3. The semiotic status of gesture

The following contextually and functionally restricted set of lexicalised gesture stereotypes will serve as an initial simplified domain: Churchill's use of the victory sign (reinterpreted in today's youth cultures as a "peace" sign); the military salute (in nation-specific variants); an air kiss to a parting close friend; a wave; beckoning (culture-specifically) with one or more fingers or the head; a dismissive hand movement; cupping a hand around the ear; the f-sign or "the finger".

All these gestures are, at first glance, simple to categorise: hand movements with some communicative function. They are all signs. But there is more to it: these gestures initiate or terminate a phatic (Malinowski 1923) interaction phase and either start or end some kind of communicative encounter, or an episode in an encounter, like the phatic interjections "Hi!", "Bye!" and like chant-like phatic "call contour" intonations (Gibbon 1976b) used in some vocative contexts ("Joooh-neee!") and with phatic interjections ("By-ye!").

The wave is a clear case of a speech-like gesture and lends itself to a standard style of dictionary definition (*definitio per genera proxima et differentia specifica*):

A *wave* is...

(1) ... a (possibly iterated) *movement* (*M*) of the hand realised with *constituent* (*C*) shape and movements of fingers, and movement of arm *environment* (*E*);
(2) ... interpreted with the phatic pragmatic meaning of initiating or terminating a dyadic discourse encounter.

There are many less clear cases, but the wave is one of the clear cases of visual-vocal gesture likeness. In fact it works like the interjection morpheme "hello" with which it may co-occur: the wave articulation...

(1) ... is holistically meaningful (*M*), like an interjection morpheme such as "hello" or "bye";
(2) ... has the same phatic pragmatic meaning as the phatic interjections 'hello' or 'bye' of initiating or terminating a dyadic discourse encounter;
(3) ... is realised with constituent phoneme-like sub-gestures (of hand, fingers, arm) which have distinctive, but not meaningful status (*C*);
(4) ... has a dynamic assimilatory effect on movements of the immediate environment, i.e. arm, shoulders, perhaps torso (*E*);
(5) ... may be repeated (iterated, reduplicated) like an interjection.

As Kendon (1996) noted:

> This "strand" of activity (which we also refer to when we use the term "gesture" or "gesticulation") has certain characteristics which distinguish it from other kinds of activity (such as practical actions, postural adjustments, orientation changes, self-manipulations, and so forth).

However, gestures do often closely resemble other kinds of practical behaviour which have no semiotic import. It is useful to distinguish three groups of such practical behaviours, with increasing intentionality and consequently potential semiotic value.

(1) *Fortuitously triggered or concomitant movements*: e.g. stumbling, stubbing the toe, banging the head, staggering. Fortuitous and concomitant behaviours are culture-independent, though ensuing gestures and interjections may well be highly culture-dependent, depending on sensitivity to pain and on community and individual conventions.

(2) *Deliberate but practised, routinised behaviour*: e.g. visible swallowing, walking, washing, scratching, basic eating and drinking. Deliberate routinised behaviour is partly culture-independent and determined largely by the human anatomy, but partly governed by conventions.
(3) *Goal-directed artefact manipulation*: e.g. moving furniture, combing the hair, applying make-up, tying shoe-laces, formal eating and drinking. Goal-directed artefact manipulation is culture-dependent and the artefacts are often highly culture-specific, e.g. eating utensils with varying gripping conventions (knives, forks spoons; chopsticks; kebab sticks; finger food).

A feature of non-semiotic but convention-constrained behaviours is that they can be used semiotically, namely when one of the conventions is deliberately or perhaps involuntarily infringed, and thereby "makes a statement" conveying a personal stance of social distance from a certain group, or insulting members of a group. A classic example (Type 2, deliberate, practised or routinised behaviour) is when someone leaves a room and bangs the door shut to mean disapproval, anger or insult (which may or – with loss of face – may not work). Such categories are parameters with values which may appear simultaneously, and are not mutually exclusive taxonomic properties: artefacts which are routinely manipulated in certain contexts – for example, tying shoe-laces – may have semiotic import in other contexts, e.g. in choosing to wear laced shoes as opposed to sandals or no footwear.

The next step along the road to complex semiotic functionality is found in other systematic behaviours which need to be integrated into a comprehensive taxonomy of gesture types (cf. Proxemic Theory, Hall 1959):

(1) *Posture*, a holistic configuration of the body, such as the so-called "Gothic s-curve" and zig-zags of Hogarth's caricatures (cf. Figure 4), or putting body weight on one leg and crossing or moving the other (German: the *Standbein–Spielbein* posture). Posture can be interpreted, for example, as deferential, threatening, sexually suggestive. Posture change is also gesture.
(2) *Orientation*, positioning the entire body to be facing, near-facing or with the back to the interlocutor. Orientation can be interpreted as attentive or as impolite. Orientation change is gesture.
(3) *Distance*, positioning the body to be far from, close to, or very close to the interlocutor (as with Vladimir Nabokov's tragi-comical Timofey Pnin and his culturally inappropriate Russian behaviour in the USA

Figure 4. William Hogarth: *Mariage à la mode, 1: Le contrat de marriage*. (Extract, brightness and contrast enhanced.)

(Nabokov 1957). Distance can be interpreted, for example, as rejection or as intimacy. Distance change is gesture.

(4) *Clothing* is communication: like dialect, sociolect, style and register in speech and text, clothing (and toying with clothing) and body decoration with make-up and jewellery have a conspicuous semiotic function of indicating membership of groups defined by regional origin, social status, formality, activity type, and individuality, or of signalling sexual attraction. Choice of clothing signals not only individual taste, but also acceptance or rejection of group norms.

Like forms of vocal and visual communication which convey pragmatic information, these communication modes can lead to positive and negative reactions, and to serious sanctions if group norms are flouted too non-conformistically: by ostracism, legal action, or violence.

Beyond the types of semiotic behaviour already discussed, there are numerous situation-conditioned gestural communication systems, including those which are an integral part of professional speech communication registers, typically in acoustically hostile environments: the communication register of a sports umpire or referee; gestures and the "Lombard Effect" in the speech of stock exchange floor dealers in London's Lombard Street, with characteristic voice quality changes when shouting; the "bat" waving of the batman directing

aircraft into parking position; semaphore signalling in earlier maritime communication. Much gestural communication is *teleglossic* (distance) communication, which ranges from waving, whistling and calling to registers used in telecommunication (Gibbon 1985; Gibbon and Kul 2006).

So far there has been little overall systematisation of the sociolinguistics of gesture in terms of classic dimensions of language variation, an open field of research (but cf. Rossini 2004):

(1) *Idiolect*: personal gestural patterns and habits, including quirks and twitches.
(2) *Region*: cultural and dialectal variation in gesture, both local and global, such as inter-communicator contact with hand-shaking, in Western Europe, as opposed to avoidance of such contact by means of a hand-on-heart gesture for greeting in Iran, or a palm-to-palm gesture for greeting in South and East Asia (a prayer gesture for Christians).
(3) *Society*: indicating socio-economic group identity, for instance in greeting gestures, behaviour at table.
(4) *Function*: gesture register based on activity and occupation, often in acoustically unfavourable environments, such as sports, stock exchange and airports, and in teleglossic communication.

The topic of language variation is closely linked to the issue of universals. The naïve view is that one's own gesture is universal, one's own prosody is universal, and – the extreme naïve view in unilingual communities – that one's own language should be universal, even if it is evidently not. But there is also much folk-lore about culture-specific differences: the manual gesture of a thumb-and-forefinger loop, for instance, has different meanings in different parts of Europe, from a simple iconic shape for a circle, "zero" or "O" through a symbolic "very good", to the extreme of an insulting iconic metaphor meaning 'cunt'. Less extreme cases are different meanings assigned to head-shaking (restricted horizontal rotation), nodding (restricted vertical rotation) and head waving (side to side movement with no rotation) in different communities. Nevertheless, a few tentative universals can be proposed:

(1) *Universals*:
    (1) All communicator communities use both vocal and non-vocal communicative gestures.
    (2) Visual gesture functionality parallels prosodic and/or lexical functionality.

(3) Functions of gesture are in general universal.

(2) *Specifics*:
   (1) Lexical gestures (see the following section) are specific to cultures and environments.
   (2) Forms of gestures (even those with universal functions) are in general culture specific.

3. Models of the functions of gesture

A selection of linguistically motivated models of the semiotic functionality of visual and vocal gesture will be outlined in this section, and explicitly related to different kinds of gesture. The wide range of dimensions involved appear *prima facie* to be intractable, but on closer inspection they lend themselves to a heuristic dimensionality reduction in the form of universal unidimensional scales such as McNeill's (1992) *Kendon continuum* (cf. Table 1; not a continuum, actually, but discretely partitioned). A more comprehensive scale is proposed here: a "Natural-Conventional Scale" (NCS) for gesture functionality (cf. Figure 5; "conversational gesture" covers both prosodic and lexical functionalities; "encoded gesture" covers "artificial gesture languages" such as semaphoring.).

Table 1. McNeill's model (1992) of Kendon's continuum.

→	→	... "continuum" ...	→	→
Gesticulation	Speech-linked gestures	Emblems	Pantomime	Sign Language
(obligatory presence)	(presence of speech)	(optional presence of speech)	"obligatory absence of speech"	"obligatory absence of speech"

The reduction of many dimensions to one inevitably results in artefacts in the NCS and in McNeill's scale, such as the appearance of a continuum when in fact the scale is categorial, not continuous (though there are fuzzy overlaps, e.g. between conversational and rhetorical gesture).

The study of gesture has always involved discourse analysis, both in the context of traditional rhetoric and conversational gesture (Kendon 2004), or in

conversation analysis, and a few of the major functional linguistic models of the 20th century will be examined for applicability to vocal and visual gesture.

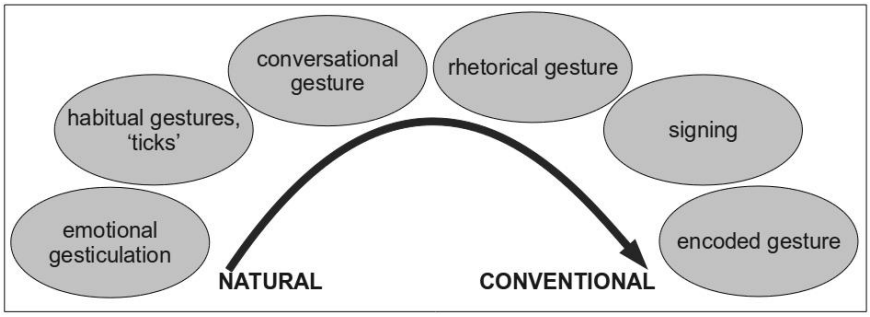

Figure 5. Natural-Conventional Scale (NCS) of gesture functionality.

3.1. Jakobson's Constitutive Factor model

The functional model of Jakobson (1960) is a configuration with six constitutive factors and six communication functions as relations between the factors. Jakobson's model extends the instrumental *organon* model of Bühler (1934), which had four factors (German: *Zeichen* 'sign'; *Sender* 'sender'; *Empfänger* 'receiver'; *Kontext* 'context'), and three functions (*Ausdrucksfunktion* 'expressive function', a relation between sign and sender; *Darstellungsfunktion* 'representation function', a relation between sign and context; *Appellfunktion* 'appeal function', a relation between sign and receiver). Bühler defines the last of these coquettishly with reference to the English collocation *sex appeal*.

Jakobson adds two constitutive factors, the *contact* (channel or medium) and the *code* (the language), renaming the sign as *message*, and defines three additional functions, the *metalingual* function between the message and the code, the *phatic* function between the message and the channel, and the *poetic* function between (parts of) the message and (other parts of) the message. In Figure 6 Jakobson's visualisation of the constitutive factors is shown, enhanced by addition of lines indicating the six communicative functions.

Jakobson's six communicative functions are well-adapted to characterising gesture functions:

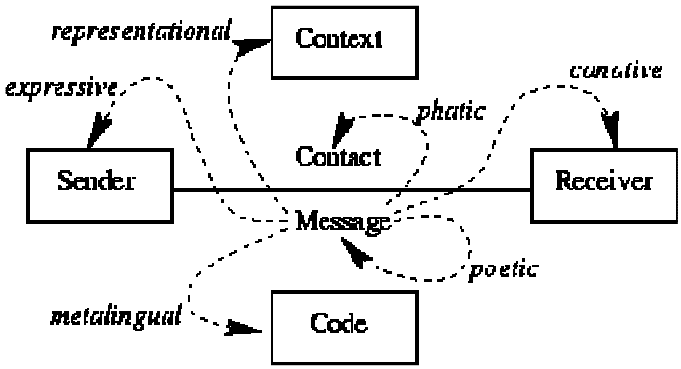

Figure 6. Jakobson's functional model (1960) of constitutive factors, with functional relations between factors added.

(1) *Expressive* gestures are not hard to find - smiles and clapping are just two of many, often accompanied by features of prosody and by locutionary expressions with the same function.
(2) *Conative* gestures aim at influencing the perceiver, e.g. beckoning with finger, hand or head, promising by handshake, and warning gestures.
(3) *Representational* gestures include deictic pointing and iconic size indication, as well as emblems and icons for drinking, eating, telephoning, writing.
(4) *Phatic* gestures, like phatic prosody (Gibbon 1976b), include previously discussed greetings and farewells, and dialogue-sustaining gestures like hand-cupped-around-ear to indicate a perception problem which hinders uptake of the utterance.
(5) *Metalingual* gestures include the beats and sweeping gestures which indicate points and intervals of particular importance in the utterance. These gestures, like accents (stresses) and emphasis particles, denote temporal locations in the utterance itself rather than in the environment, thus taking on metalocutionary, in particular metadeictic functionality (Gibbon 1983).
(6) *Poetic* gestures are used in an extremely wide range of contexts, and include not only gestures of graceful gesticulation or song and rhyme accompanying gestures but also the intricate and highly conventionalised gestures and postures of dance.

3.2. Speech Act Theory

A functional model which has already been applied to gesture modelling (Gibbon 2005) is the speech act model (Austin 1962; Searle 1969) of the *locutionary*, *illocutionary* and *perlocutionary* functions of speech acts, shown in the following examples:

(1) Deictic gestures of location and direction, as well as those indicating size and shape, for example, are primarily *locutionary*, with conventional denotational semantics, as are lexical "emblem" gestures for specific objects and activities.
(2) A nod or shake of the head, meaning 'yes' or 'no', respectively, is primarily *illocutionary*, indicating the conventional dialogue status of an utterance, in this case agreement or disagreement.
(3) A head-tapping or finger-protruding gestural insult, or a thumb-raising or clapping gesture is primarily *perlocutionary*, being intended to have a specific direct negative or positive effect on the addressee (an intention which may or may not be conveyed, or sometimes an effect which may not be intended).

The speech act model permits gestures to be polysemous. For example, a lexical gesture like "thumbs up" in a West European context may have multiply polysemous meanings of all three kinds: a locutionary meaning of success, an illocutionary meaning of agreement, and a perlocutionary meaning of encouragement or, in other contexts an insulting iconic ithyphallic connotation of 'dick' or 'prick'.

Searle's conditions on felicitous speech acts are correspondingly straightforward to apply to gestures. A particularly clear case is *Condition 1*, that "normal input and output conditions obtain". *Condition 1* specifies the phatic function of greeting and farewell gestures, intonations and interjections which have already been detailed: determining "normal input and output conditions" is a prerequisite for successful dialogue. Another clear case is Searle's *Condition 6*, the sincerity condition, evidently contravened by "magical" (Malinowski 1923) gestures with lie-licensing function, such as finger-crossing behind one's back in the schoolboy culture of my childhood (Opie and Opie 1959). Oddly, many quite normal semiotic functions of bewitching, deceiving and annoying do not figure much in the standard literature, which apparently prefers to teach prescriptively about "nice" and "acceptable" communicative behaviour.

3.3. Grice's Maxims of Cooperation

Another speech act model which applies directly to gestural communication is Grice's (1989) functional model of Maxims of Cooperation, which are descriptive, not prescriptive and essentially semantic (though often called "pragmatic"):

(1) *Maxim of Quality* (be truthful): this applies to representational (e.g. emblematic, iconic, deictic) gestures, as well as gestures of agreement and disagreement.
(2) *Maxim of Quantity* (be as informative as and not more informative than required): differences in the understanding of what gesture "quantity" means vary greatly from culture to culture, and also individual to individual: Italians are stereotypically said to gesture more than Brits.
(3) *Maxim of Relation* (be relevant): gestures should have a bearing on the current interaction, and not be arbitrary (cf. also the Maxim of Quality).
(4) *Maxim of Manner* (be clear): gestures should be clearly distinguished and not ambiguous.

Not all communication is cooperative. Aggressively expressive or deceitful "double-bind" prosody and gesturing are not cooperative, but involve contraventions of the maxims. A non-committal smile violates the Maxim of Manner and the "magical" gesture of crossing the index and middle finger behind one's back makes it okay to lie, contravening the first maxim. Non-Gricean behaviour is at least as "normal" as Gricean behaviour. But many gestures, like prosodic patterns, are highly routinised and used subconsciously, which, perhaps fortunately, makes them difficult to use uncoooperatively.

3.4. Leech's Maxims of Politeness

Unlike Grice's semantic maxims, Leech's maxims (Leech 1983) are pragmatic in that they concern social relations, and relate easily to gestures:

(1) *The Tact Maxim* (minimise cost to the other): gestures of acknowledgment, or looking away if a mishap occurs.
(2) *The Generosity Maxim* (maximise generosity to the other): gestures which foreground the other, such as smiling or waving someone forward.

(3) *The Approbation Maxim* (maximise approval of the other): the thumbs-up gesture which has already been mentioned, and other gestures of encouragement and approval, such as hand-clapping.
(4) *The Modesty Maxim* (minimise self-praise): a gesture of self-deprecation such as a gentle waving movement with palms downward.
(5) *The Agreement Maxim* (maximise expression of agreement with the other): back-channel gesture articulations signalling agreement, disagreement, encouragement, turn-yielding.
(6) *The Sympathy Maxim* (minimise antipathy to the other): from handshaking, back-clapping, hugging, to various kinds of kissing, distant or passionate, and other intimate behaviour.

Like Searle's felicity conditions and Grice's maxims, these maxims of politeness are not prescriptions of "good behaviour" for "nice people", but characterisations of a culture-specific behavioural space of friendliness and hostility, intimacy and distance, along the six dimensions. Behaviour which infringes the pragmatic maxims of politeness can be designated non-Leechian, by analogy with the non-Gricean behaviour defined as infringements of the Maxims of Co-operation.

3.5. McNeill's gesture taxonomy

The most well-known taxonomy of gesture functions is probably that of McNeill (1992), which is quite closely related to the functions described in the earlier linguistic models of communicative functions, as well as to categories from traditional Peircean semiotic theory (Peirce 1958–1960). The range of functions also relates seamlessly to the linguistic taxonomies discussed previously. The overview in Table 2 shows a selection of the functions discussed by McNeill and associates in various publications (cf. 1992, 2005, and references there) and related prosodic and lexical properties of speech.

In relation to the McNeill taxonomy, the phatic greeting and farewell gestures which have already been introduced again form a particularly interesting case of items which need further explanation. Phatic gestures are emblems, resemble interjections, may also have indexical (deictic) function, and relate not only to the prosody of call contours, but also to iconic surrogates of prosody such as whistling. The phatic gestures resemble interjections in that they have a relatively fixed but often hard to define form-function relationship, and they are extra-grammatical, i.e. they do not fit into the regular flow of speech but have

an autonomous attention-getting, channel-creating or emotional status. The same conditions apply to the chant-like stylised phatic intonation (Gibbon 1976a, 1976b) used in calling, routine lists and corrections, and with some interjection-like greetings ("Hello-o!") and farewells ("By-ye!").

Other emblems, icons and deictics are more clearly related to the main parts of speech in language. Like other parts of speech, their forms are highly language or culture specific.

Table 2. Correspondence of gesture characterisation with components of speech.

Gesture type	Characterisation	Example	Speech correspondences
Emblems	Fairly highly conventionalised, lexicalised gestures for culture-specific common activities, constituting the most well-known tye of gesture. Emblems correspond to Peircean *symbols*.	Phatic greeting/ farewell (hand-shaking, waving), insulting (tongue-protrusion, finger-protrusion); activities (phoning, writing; eating); attitudes (success, pleasure; cuckoldry).	Lexicon: – Lexical words (e.g. interjections, nouns, adjectives, verbs, adverbs of manner). – Established or ad hoc (nonce) coinages. Prosody: Cf. chant-like phatic intonation.
Iconics	Resembling the referent in shape, size or manner of movement, describing an object with the hands, or transduced into an onomatopoeic sound. Iconics correspond to Peircean *icons*.	Hands high, wide apart) or manner of movement (fast, slow, iterative); onomatopoeic lip-smacking, clapping, finger-snipping, stamping.	
Metaphorics	Vehicle (the gesture) relates to tenor (non-literal meaning) of the metaphor. Two layers of semantic interpretation, literal and non-literal vehicle as emblem or icon. Multiply categorised in Peircean terms.	Application of emblems and iconics to abstract concepts, e.g. payment gesture meaning negative consequences; indicating a container or conduit for ideas, or a gift of an idea or suggestion.	[Both lexicon and prosody correspondence types apply to emblems, iconics and metaphorics.]

Deictics	May indicate an actual physical position, size, distance or direction, but may also place concepts metaphorically in physical gesture space. Deictics correspond to a kind of *index*, in Peircean terms.	Culture-specific pointing gestures with hand, index finger, head, pursed mouth, gaze.	Lexicon: – Demonstrative pronouns, adverbs Prosody: – Accentuation, emphasis
Beats	Moving roughly in synchrony with rhythm of speech, marking a sequence or hiatus, e.g. change of theme or focus. Beats are metadeictic, Peircean *indices* marking temporal locations in speech.	Regularly iterated arm and finger gesture, nodding, eyebrow-raising; gesture amplitudes indicating different status of any accompanying speech unit.	Prosody: – Beats corresponding to speech timing units syllable or foot, and as metadeictic, rhythmic and emphatic accent. – Cohesive gestures corresponding to global intonation contours and rhythms. [Both lexicon and prosody correspondence types apply to beats, butterworths and cohesives.]
Butterworths	Co-occur with disfunctionalities in speech. In Peircean terms, Butterworths are a kind of metadeictic *index*, marking temporal locations in speech itself.	Hand-waving, lip-pursing.	
Cohesives	Creating a gestalt coextensive with a spoken utterance or its parts. In Peircean terms, cohesives are metalocutionary *icons* mirroring speech structure, and metadeictic *indices*, marking regions of speech.	Slow sweeping gestures of arms, head, posture changes, indicating the extent of any accompanying speech unit.	
Affectives	Displaying emotional states and events. In Peircean terms, also a kind of *index*.	Amplitude of gestures, as well as emblems with lexicalised affective meanings such as smiling, frowning, clapping.	Paralinguistic features: Pitch height and range; tempo changes; intensity and voice quality changes.

Like the other functional models discussed so far, McNeill's function inventory is a set of parameters or dimensions in a semiotic quality space rather than a taxonomy of mutually exclusive categories, and values of the parameters can co-occur in any given gesture: emblems can be simultaneously iconic, metaphorical, emblematic and deictic. For example: holding the hands wide apart may indicate the great importance of some issue, perhaps changing in configuration as the accompanying locution changes, thereby also functioning simultaneously as a cohesive.

3.6. Interactive dialogue models

The taxonomies discussed so far have been to greater or lesser extent non-dynamic and focussed on individual contributions to interaction rather than on interactive temporal sequencing. Interactive dialogue theoretic models from interpretative conversation analysis and ethnomethodology, and from more recent human-computer interface models, also apply to gestures, however. The specific points to be demonstrated are: *turn-taking*, *back-channelling*, and *phatic*, *ludic* and *magical* interactive dialogue acts.

Turn-taking functions of gesture, gaze and posture are highly culture-specific. Gaze, for example, is strictly conventionalised, as "Look at me when I'm talking to you!" or "How dare you look at me like that!". For example: in Western Europe a common dialogue convention is for the addressee to look at the speaker, but for the speaker's gaze to wander, returning to the addressee for emphasis and turn-closing. Elsewhere it may be impolite to look a social superior in the eye. A very comprehensive taxonomy of turn management dialogue acts has been prepared by Bunt (2010); a selection Bunt's categories is shown in the overview in Table 3.

Gestures also have *back-channel* functions, as in dialogue feedback with shoulder-shrugging, eyebrow-raising, nodding and head-shaking, smiling or mouth-corner depression for various shades of approval or disapproval.

Phatic functions are to be found not only in the previously discussed cases of greeting and farewell, but also in hugs with *bisous* or air kisses, cheek-kisses, blown kisses, or in hand-shaking and back-slapping in more extrovert cultures,

contrasting with head inclination, bowing and prostration in more restrained cultures which proscribe public body contact between interlocutors. Another phatic function of gesture is gesture harmony (gesture-sharing, mimicry), i.e. the generally subconscious replication of gestures and postures by interlocutors, indicating mutual rapport and bonding. Gesture harmony is closely related to

Table 3. Adapted Dynamic Interpretation Theory categories (DIT; Bunt 2010) applied to gesture.

Dialogue act functions	Gesture examples
1. General Purpose communicative functions	
1. Information transfer	
1. Information seeking	Querying gestures (e.g. raised eyebrows)
2. Information providing	Raised or wagging ("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

the frequently observed adaptation of prosody (rhythm and melody) in conversation, for instance in high-pitched intonations and low-amplitude concomitant gestures of the baby-talk register.

Ludic functions of gesture in dialogue are common: "Gimme five!", "Hi five!", partly as greeting, partly as agreement or solidarity, partly as play. Ludic functions are also found in play and in gesture accompaniments to songs and

rhymes. A well-known English iconic example, is "Incey Wincey spider", deriving from the Akan name "Anansi", the wise spider of stories in West Africa and the Caribbean, with spider-like ambidextrous fingertip-touching climbing movements. In the USA, a phonetic change to "Itsy bitsy spider" developed and a semantic change occurred, from a name to an adjective meaning 'very small'.

Magical (Malinowski 1923) functions of gesture in dialogue are comparable to incantations such as "abracadabra". They are found in "good luck" gestures like finger-crossing or "touch wood" among English speakers, or thumb-squeezing (*Daumendrücken*) among German speakers, and are related to religious gestures such as the Christian "sign of the cross" as a sign of reverence or invocation of divine support, and other liturgical gestures and postures in different religions.

4. Models of the forms of gesture

Vocal and visual gesture articulations require formal modelling for a full understanding of the dynamic processes of communication. This applies both to theoretical uses for hypothesis consistency checking and analysis of large quantities of data, and for practical uses in domains ranging from language teaching or diagnosis and therapy of behavioural problems to the development of software for robotic and video applications. Formal models of the compositionality of vocal and visual articulations will be proposed.

4.1. From rhetoric to robots: gesture compositionality

The study of the forms, as well as the functions of gesture, goes back to the study of rhetoric in ancient Greece and Rome, to ancient holy scriptures, for instance, and extends to the thespian lore of gesture in dramatic acting (see Kendon 2004 for a rather comprehensive overview). Formal models for gesture articulations are mainly to be found in artificial intelligence and robotics, where complex forms are (at least initially) are more relevant than complex functions, non-semiotic movements such as grasping and turning are more relevant than semiotic gestures, and semiotic gestures (with the exception of sophisticated video games) are often reduced to sets of deictic stereotypes for denoting positions in space-time, or to naïve one-to-one emotion-gesture stereotypes. Relatively recently the term "gesture" has also been associated with computer input devices: finger movements on a touchpad; arm and wrist movements directing a

computer mouse in two dimensions; button clicking; wheel turning; joystick control; body-movement detectors for video sports games; acceleration and orientation detectors on smartphones.

In an influential overview paper, Kendon (1996; cf. also Kendon 2004) discusses a range of previous approaches and proposes an agenda for future studies. To some extent, this agenda is already being fulfilled by a host of different practical interests, from sign languages of the hearing-impaired to software development for robotic systems and video games. These recent developments require operational models, i.e. models with dynamic temporal behaviour, over and above the previously discussed function taxonomies. In the following discussion, the required dimensions of *paradigmatic*, *syntagmatic* and *realisational* relations will be discussed.

4.2. Paradigmatic relations

Paradigmatic relations of similarity and difference are traditionally dealt with in taxonomies, an example of which is a thesaurus. In discussions influenced by Artificial Intelligence, the term "ontology" (originally the study of what *exists*, the science of *being*) was first introduced for taxonomies ("is-a" relations, including "folk taxonomies" or "folksonomies'), and then extended to mereonomies (hierarchies of parts and wholes, also "meronomies", "partonomies"). A comprehensive guide to such relations in lexical semantics is provided by Cruse (1987). A proposal for linguistics itself has been developed by Farrar and Langendoen (2003) with the *General Ontology for Language Description* (*GOLD*), and an approach to formulating conditions on an ontology for prosody was formulated by Gibbon (2009). The more general field of multimodal communication remains almost virginal in this respect (but cf. Gibbon et al. 2000; Gibbon 2005).

The paradigmatic properties of items in taxonomies are often formalised as feature structures (attribute–value pairs and attribute–value matrices), plus constraints on combinations of properties. These feature structures adorn the nodes of classificatory trees in many modern models of language structure and figure as feature bundles in phonology or as attribute–value matrices in formal syntax and semantics. Paradigmatic relations of similarity and difference may pertain to physical form (e.g. gesture patterning in visual and vocal communication), to structure (similarity and difference in syntax), and to meaning (semantic and pragmatic similarity and difference). A fully formal account of paradigmatic relations in visual and vocal communication, involving all possible modalities,

Table 4. Partial matrix for characterising a taxonomy and quality space of functional gesture types mapped to coarse-grained specifications of gesture articulators (visual; head and upper body only; posture and distance excluded).

Articulators (body areas)				Functional multimodal complexes							
				Phatic			Speech act			Appraisive	
				Wave	Salute	Air kiss	Agree	Don't know	Obliga-tion	F-sign	Thumbs-up
Head	Skull						nod				
	Face	Eyebrows						raise			
		Eyes									
		Nose									
		Mouth	Lips			purse					
			Tongue								
			Velum								
			Larynx								
Upper	Shoulders							shrug			
	Arms			raise	raise & bend						
	Hands			flap					clasp	palm inward	
	Fingers				extend					spread i & m	
Orientation				facing	facing	facing					
Participants				>1	>1	1	1	1	2	>1	>1

has not yet been attempted, and is far beyond the scope of the present contribution.

A selection of visual modality complexes (sets of pairs of motor articulators and the visual sense organ) is illustrated in Table 4 in a sparsely populated matrix. Greater granularity of detail for visual and vocal articulators is needed for full specification. The articulators and their functions are represented as two hierarchies of attributes, with sample values in the cells of the matrix. And, of course, gestures do not only generate visual output, so specifications of perception are also needed:

(1) *Visual*: the gestures categorised in Table 4.
(2) *Acoustic*: speech; finger snipping, clapping, stamping, as well as speech surrogates such as whistling, drumming. Music also falls into this general category.
(3) *Haptic* (*tactile*): hand-shaking, back-slapping; hugging, kissing, caressing; erotic contact.
(4) *Olfactory*: voluntary or involuntary smell (scent; gift of flowers; food, drink; physical proximity).
(5) *Gustatory*: voluntary or involuntary taste (food, drink; consequence of erotic contact).

4.3. Syntagmatic and realisational relations: The mereonomy of gesture

Studies of gesture grammar have produced inventories of hand configurations (Martell 2001) and of linear *preparation–hold–stroke–hold–retraction* patterns, sometimes with finite-depth hierarchies (Kendon 1972; Gibbon et al. 2003). The following formulations nicely summarise the formal syntagmatic features of a gesture articulation as a "phrase of action" metaphor with linear structure (Kendon 1996):

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

The four characteristics which need to be noted are: movement is away from and back to a rest position; the movement has a "peak", "centre" or "stroke"; the temporal boundaries of the movement are clearly demarcated (unlike gradual changes in orientation or posture); the movement is symmetrical, in contrast to

many other actions (reverse spooled video may be difficult to distinguish from the forward spooled video).

Kendon's description is strongly idealised, but harmonises in general with principles of articulatory phonetics, with the restriction that both vocal and visual articulations may assimilate to adjacent articulations and therefore may not reach target or rest positions. As in speech, it is likely that gestures have no linear temporal asymmetry, but that patterns have a logarithmic temporal structure. Three syntagmatic properties have been noted so far which must be accounted for: linear basic structure, iterations, finite-depth hierarchies. From a formal point of view, such properties indicate that "linear" or "regular" grammars (and finite state machine models for these grammars) will be adequate.

The three properties may be necessary conditions on gesture structures, but they are not sufficient conditions. If there is recursivity beyond iteration in gesture patterns (Rossini, in press) then more complex grammars are needed. Also, the internal composition of gestures requires additional feature models. Finally, temporal relations between articulator events also require different additional structure and real-time temporal properties of gestures must be considered.

In the following discussion, these syntagmatic relations between articulator events will be discussed in terms of a *Linear-Feature-Timing-Realtime* (*LFTR*) model, building on insights from the study of vocal gesture in speech:

(1) *Linear Precedence* (*L*): models temporal event sequences and hierarchies.
(2) *Feature Structure* (*F*): models internal structure of events in terms of feature combinations, the values of which define similarities and differences.
(3) *Timing Relation* (*T*): defines temporal relations of precedence and overlap of events independently of actual time measurements; sometimes referred to as "rubber time".
(4) *Real-Time Interpretation* (*R*): defines measured durations of the intervals of events and between events; sometimes referred to as "clock time" (subjective estimations of interval durations are sometimes called "cloud time").

The *LFTR* model is needed for modelling the prosody and phonology of vocal gesture articulation, and likewise applies to the "prosody" and "phonology" of visual gesture articulation.

4.3.1. The *Linear* (*L*) component

The *Linear L* component is the starting point for compositional systematisation of the parts and combinations of gesture articulations and coarticulations (the modification of articulations by adjacent articulations).

Kendon's previously cited definition of "gesture" as "clearly demarcated symmetrical movement from a rest position via a peak (centre or stroke) back to a rest position" refers primarily to the linear composition of "atomic gestures", which are the basic "morphs", i.e. smallest meaningful segments, in vocal and visual articulation streams. The structure of an atomic visual gesture articulation resembles the sonority curve of prototypical CVC syllables in speech from rest through a low sonority initial consonant and a high sonority vocalic segment to a low sonority final consonant and rest. Alternatively, the articulation of any segment follows the same principle. But the "syllable" analogy, at least, is a little misleading: visual gestures have meanings, syllables do not, so visual gestures are more like monosyllabic morphemes than syllables. The smaller components of iterated gestures such as waving are like morphemes in sequence. Other complex gestures such as the "sign of the cross" are more like disyllabic simplex words, combining vertical and a horizontal "syllable" gestures. The morpheme-like gesture contrasts with a "gesture compound word" such as indicating size with two hand gestures in a complex gesture, and this contrasts with the "gesture phrase", i.e. a syntagmatic combination of two independent 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!'. Sign languages use the same syntagmatic phrase principle in much more complex ways, like speech.

Operational computational "working" models of gestures have been developed both for theoretical reasons in hypothesis checking and for practical purposes in robotics and avatar development. Each development has necessarily had to deal with gesture syntax. Several proposals have been made, two of which will be briefly characterised: *CoGesT* ("Conversational Gesture Transcription") and *MURML* ("Multimodal Utterance Representation Markup Language"). The basic *Source–Stroke–Target* articulation structure is represented differently in the two approaches: in *CoGesT*, articulation structure is represented directly by a *Source–Trajectory–Target* triple, while *MURML* uses a *Source–Direction–Distance* format which is more convenient for calculations in a robotics application environment, but does not address global trajectory shape.

MURML is a set of XML conventions (Wachsmuth and Kopp 2002) for representing gestures in a robotics context with conversational agents. Features

used are: *Timeline*, *Symmetry*, *HandShape*, *PalmOrientation*, *ExtendedFingerOrientation*, *HandLocation*, *ShoulderLocation*, *CentreLocation*, *Start*, *Direction*, *Distance*. The *MURML* specification has very many more details for attributes of the hand than can be discussed here.

The linguistically motivated *CoGesT* model (Gibbon et al. 2004) concentrates on syntagmatic precedence and overlap operations and provides a formal grammar (for hand gestures only). Figure 7 contains three frames from a recording of a German narrative by a professional story-teller, and shows a *CoGesT* analysis of a bimanual appellative gesture.

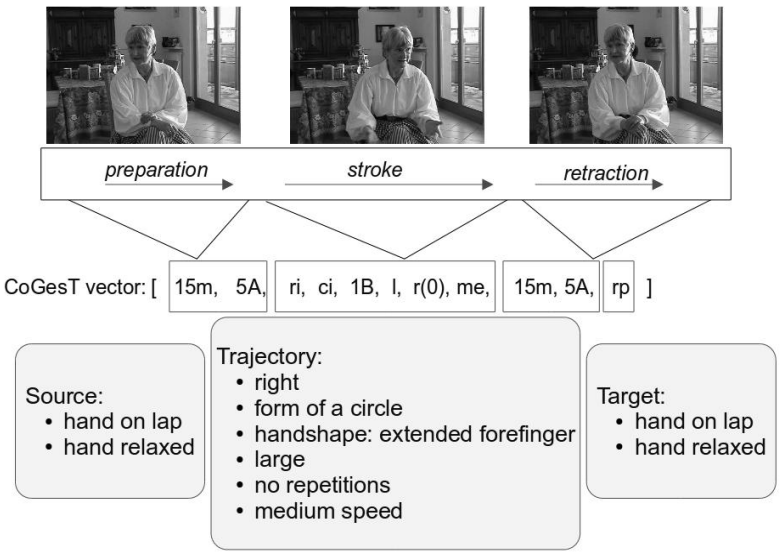

Figure 7. Application of CoGesT transcription to an iconic gesture.

Atomic or simplex gestures of the kind shown in Figure 7 are of two types: *2-place static*, i.e. hand *Shape* and *Position*, and *9-place dynamic*, i.e. *Source* (*Location* and *Handshape*), *Trajectory* (*Lateral*, *Sagittal* and *Vertical Direction*; *Shape*, *Form*, *Size* and *Speed*), *Target* (*Location* and *Handshape*). These attributes are represented as a vector or feature structure optionally enhanced with specifications for two-member gestures, *symmetric* (where hands make mirror

image movements) or *parallel* (where hands make the same movement) and with indicators for left or right side of the body with paired articulators, here left and right hand. The model also has a specification for iterative articulations, e.g. beats, waving.

For the *CoGesT* analysis of syntagmatic gesture relations into components, a formal gesture grammar was designed. A formal grammar is not an end in itself, but an instrument for avoiding misunderstandings. For computational applications in gesture synthesis for avatars, robots or video games, a formal grammar is obviously a requirement. The following rule-set defines the *CoGesT* formal grammar (cf. Gibbon et al. 2003 for further detail) in a standard notation for context-free (Type 2) grammars:

```
<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 *CoGesT* context free syntax expresses hierarchical syntagmatic relations of the kind found in grammars for word and sentence structure, albeit in this case

of finite depth (bar iteration). The distinction between sequential and simultaneous compositionality is not explicitly modelled but is dealt with in the F and R components of the *LFTR* model and must be specified for each rule.

4.3.2. The *Feature Structure* (*F*) component

Features and feature structures were introduced in connection with paradigmatic relations and will therefore be mentioned only briefly here (cf. Bressem and Ladewig 2008 for a feature model of hand gestures). Attribute-value structures have both paradigmatic and syntagmatic dimensions:

(1) The value set of an attribute (e.g. the actual values in a *CoGesT* feature vector) represent paradigmatic relations.
(2) An attribute-value structures as a whole (e.g. a *CoGesT* feature vector) represents a syntagmatic relation of compositionality.

The use of attribute-value sets in these way is illustrated in the *CoGesT* analysis in Figure 7.

4.3.3. The *Timing Relation* (*T*) component

Atomic vocal and visual gestures enter into complex *Timing Relations* with other atomic gestures; this is particularly obvious in the study of speech prosody: intonations, like gestures, are temporally parallel to locutions. In the bimanual gesture shown in Figure 7, two atomic gestures are synchronised into a complex bimanual gesture. The "wave and smile" combination is a familiar stereotype.

The *Time Type Model* (Gibbon 1992, 2006) provides a suitably detailed framework for vocal and visual articulation modelling by abstracting three temporal levels of description: *Categorial Time*, *Relational Time* and *Absolute Time*:

(1) *Categorial Time* (functional, abstract time): abstract linguistic property or category, e.g. of [±duration], for significant gesture durations, phoneme length contrasts etc., or of a concatenation operation modelling spatio-temporal left–right/before–after articulation sequences. Descriptions in *Categorial Time* enter into realisation relations with descriptions in *Relational Time*.
(2) *Relational Time* (functional, "rubber time"): precedence and overlap relations formalised, for example, in Event Logic (van Benthem 1983)

and Interval Calculus (Allen 1983). Descriptions in *Relational Time* enter into realisation relations with descriptions in *Absolute Time*.

(3) *Absolute Time* (physical, "clock time"): a vector of measurements, as in recordings and annotations of digitised audio and video signals. Descriptions in *Absolute Time* are dealt with in the *Real-time Component* of the *LFTR* model.

A further time type may be defined, which is often found in informal transcriptions of durations, namely *Subjective Time*, "cloud time".

Allen's Interval Calculus defines the 13 (or 14, counting equality symmetry) possible relations between two intervals (Figure 8), e.g. between visual and vocal articulations. An *articulation event* is a pair of *movement* and *interval*:

ARTICULATION=<*MOVEMENT*,*INTERVAL*>

Synchronisation of events is an Allen Relation between articulation intervals:

ALLEN(*ARTICULATION*$_X$,*ARTICULATION*$_Y$)

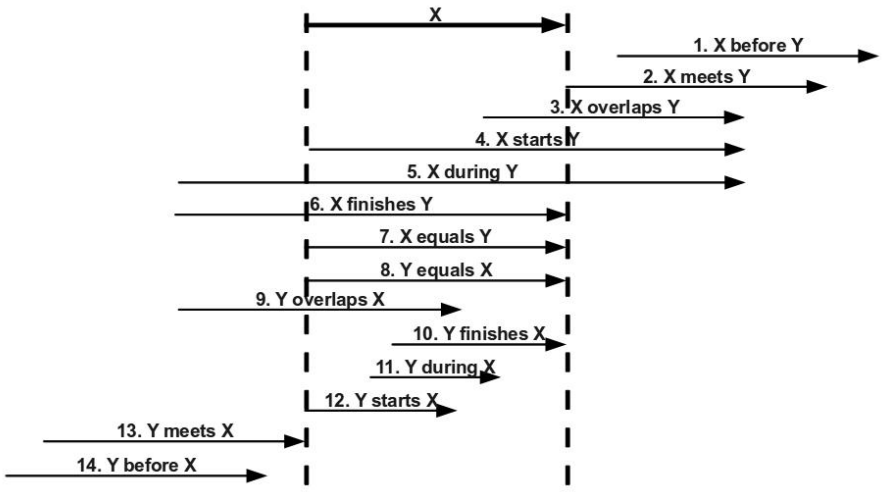

Figure 8. Interval relations in Allen's Interval Calculus.

Carson-Berndsen (1998) has shown how interval and event structures of the kind shown in Figure 8 can be formalised within the Time Type framework as finite state transducers which map between Time Types. Thies (2003) has confirmed empirically that for certain types of gesture there is a displacement relation: the delayed synchronisation relation between a hand gesture and an associated word constituent is typically, in terms of Allen interval relations, one of the following:

$$OVERLAPS(ARTICULATION_{HAND},ARTICULATION_{WORD}) \text{ or}$$
$$BEFORE(ARTICULATION_{HAND},ARTICULATION_{WORD}) \text{ or}$$
$$MEETS(ARTICULATION_{HAND},ARTICULATION_{WORD})$$

The Allen model for this "hand-mouth" relation does what a good model should: it suggests new hypotheses. For speech prosody, it has been verified in perceptual experiments that different accent synchronisation types are significant (Kohler 1987) and it is not unlikely that this also applies to visual gesture articulations, though "hand–mouth" timing is not as fine-grained as speech timing.

4.3.4. The Real-Time (R) component

The *L*, *F*, and *T* components of the *LFTR* model cover necessary but not sufficient properties of inter-articulation relations. Physically measurable time must also be considered. The *Relative Timing* modelled by the Allen relations needs further interpretation in terms of *Absolute Timing*: the "*BEFORE*" relation gives no information about whether the intervening interval is 300 milliseconds, seconds, hours or years. Physical "clock time" information is provided by the *Real-Time* component. In speech, this component corresponds to phonetics. An illustration of the *R* component was already given in the first section in the "phonetic" analysis of Ega story-telling (Figure 2, Figure 3).

What the models discussed so far do not represent very well is the dynamic temporal character of visual and vocal gestural communication in terms of events, i.e. processes which "happen" in time. The models which come closest to this would be those which refer to boundaries and trajectory peaks (strokes) in terms of *Time Relations*. But in order to capture the dynamism of gesture articulations as events in time, a higher degree of precision is required, plus an interpretation in terms of well-defined processes with real-time properties, i.e. an operational or "working" model of gesture.

Figure 9 shows a frame from the scene shown in Figure 7, and in the automatic reconstruction by an avatar from the *CoGesT* transcription of that scene (Trippel et al. 2004). The original video recording (left) is emulated on a computer screen by the avatar implementation (right). The transcription was transformed into to on-screen arm movements, demonstrating the correctness of the transcription in terms of the specified requirements. Inspection shows immediately that, like all models, the avatar model also contains artefacts, but the essential features of the articulation process are clearly reproduced.

Figure 9. Original video frame and avatar of synthesised iconic gesture.

The operational system demonstrated in Figure 9 points to directions of current and future research on computational reconstruction of visual and vocal gesture production, involving not only the articulatory models commonly found in some techniques for speech and gesture avatar synthesis but also more comprehensive systems of gesture production such as that outlined by Rossini (in press), based on the speech production model of Levelt (1989). And at some point, models of gesture perception will surely also appear.

5. Conclusion: The *Rank Interpretation Model* (*RIM*) of multimodal communication

The preceding discussion has explored "gesture as a linguistic domain" with descriptive and computational linguistic models of functions and forms of vocal

and visual gesture articulations in communication. The overall perspective has so far been more eclectic than integrated or unified, though many details have been covered. However, the issue of how different articulations in different modalities relate to each other in the linguistic "overall scheme of things" has so far remained open.

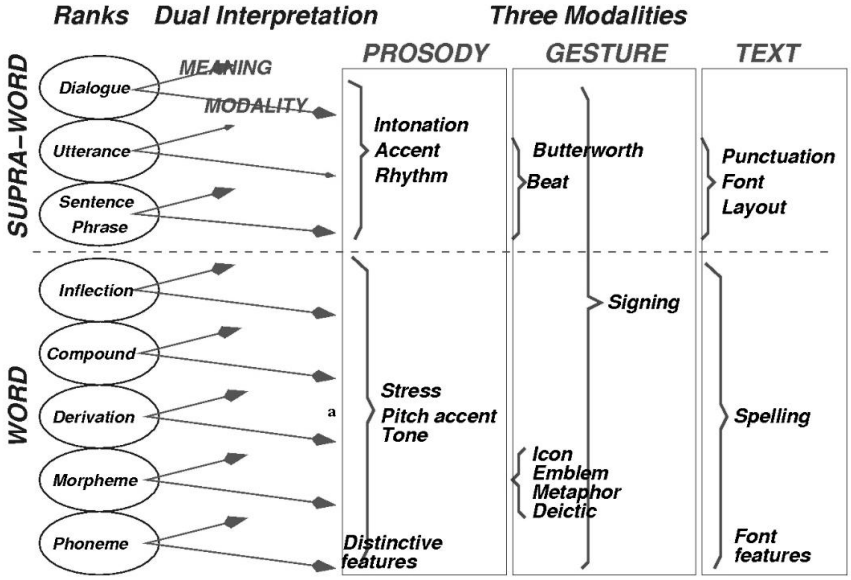

Figure 10. Rank Interpretation Model (RIM) for speech (including prosody), gesture (with lexical and prosodic functionalities) and text.

The picture is not complete without the structural and functional organisation of articulations in different modalities into ranks of different formal and functional scope. Units of speech are organised into functional ranks, from phonemes through morphemes, words (simple, derived and compound), phrases and sentences into turns and texts. This functional rank structure applies not only to vocal but also to visual gesture articulations, the visual organisation of text. The *Rank Interpretation Model* (*RIM*, outlined in Figure 10) is introduced in order to provide a more comprehensive picture of ranks and their interpretations in multimodal contexts.

The *RIM* is a 3-dimensional model relating communicative signs as abstract units to each other along a *Rank* dimension of structural and functional signs of different types and sizes, and relating these signs to an *Interpretation* of their meanings at each rank, and to a physical *Interpretation* of the articulations associated with each rank. Only the physical *Interpretation* in different modalities (more precisely: submodalities) of prosody, gesture and text are shown. Items in different modalities may well have somewhat different semiotic status, but the general structural and functional ranks are valid, as the preceding discussion of functional models has shown.

In conclusion, the present research has shown in the context of the "prosody : gesture" relation and a "lexicon : gesture" relation that gestural studies have a much closer affinity to linguistic models than has previously been claimed, and that gaps remain in the areas overed by mainstream gesture studies, some of which can be filled by descriptive and computational linguistic models. The development of integrated descriptive and computational approaches to vocal and visual gesture in communication, and to formal and operational models of these, bringing together a range of disciplines from a variety of areas, is clearly a difficult and, politically, a not uncontroversial path to take. But it is a very promising avenue for future research in theoretical and applied linguistics, both in the theory of multimodal communication and in the traditional practical application domains of therapy and teaching, as well in as the newer domains of multimodal speech technology, robotics and video simulation.

REFERENCES

Allen, J.F. 1983. "Maintaining knowledge about temporal intervals". *Communications of the ACM*, 26 November 1983. ACM Press. 832–843.

Arbib, M.A. 2006. *Action to language via the mirror neuron system*. Cambridge: Cambridge University Press.

Austin, J.L. 1962. *How to do things with words*. London: Oxford University Press.

Benthem, J.F.A.K. van. 1983. *The logic of time*. Dordrecht: D. Reidel Publishing Company.

Birdwhistell, R. 1970. *Kinesics and context*. Philadelphia: University of Pennsylvania Press.

Brentari, D. 1998. *A prosodic model of sign language phonology*. Cambridge, MA: MIT Press.

Bressem, J. and S.H. Ladewig. 2011. "Rethinking gesture phases: Articulatory features of gestural movement?" *Semiotica* 184(1/4). 53–91.

Browman, C. and L. Goldstein. 1992. "Articulatory Phonology: An overview". *Phonetica* 49. 15–180.

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

Bunt, H. 2010. "DIT++ taxonomy of dialogue acts". <http://dit.uvt.nl> Last accessed on 14 Jun 2011.

Carson-Berndsen, J. 1998. *Time map phonology: Finite state models and event logics in speech recognition*. New York: Kluwer Academic Publishers.

Clements, G.N. 1985. "The geometry of phonological features". In: Ewen, E.C. and J. Anderson (eds.), *Phonology Yearbook* 2. 225–252.

Cruse, D.A. 1987. *Lexical semantics*. Cambridge: Cambridge University Press.

Cummins, F. and R.F. Port. 1996. "Rhythmic commonalities between hand gestures and speech". *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates. 415–419.

Farrar, S. and D.T. Langendoen. 2003. "A linguistic ontology for the Semantic Web". *GLOT International* 7. 97–100.

Feyereisen, P., M. van den Wiele and F. Dubois. 1988. "The meaning of gestures: What can be understood without speech?" *Cahiers de Psychologie Cognitive / European Bulletin of Cognitive Psychology* 8. 3–25.

Gibbon, D. 1976a. *Perspectives of intonation analysis*. Bern: Lang.

Gibbon, D. 1976b. "Performatory categories in contrastive intonation analysis". In: Chiţoran, D. (ed.), *Second International Conference of Contrastive Linguistic Projects*. Bucharest, Bucharest University Press. 145–156.

Gibbon, D. 1983. "Intonation in context. An essay on metalocutionary deixis". In: Rauh, G. (ed.), *Essays on deixis*. Tübingen: Narr Verlag. 195–218.

Gibbon, D. 1985. "Context and variation in two-way radio discourse". In: Ferguson, C.A. (ed.), *Discourse processes* 8(4). 391–420.

Gibbon, D. 1992. "Prosody, time types and linguistic design factors in spoken language system architectures". In: Görz, G. (ed.), *KONVENS '92*. Berlin: Springer. 90–99.

Gibbon, D. 2005. "Prerequisites for a multimodal semantics of gesture and prosody". In: Bunt, H. (ed.), *Proceedings of the International Workshop on Computational Semantics*. *IWCS* 6.

Gibbon, D. 2006. "Time types and time trees: Prosodic mining and alignment of temporally annotated data". In: Sudhoff, S., D. Lenertová, R. Meyer and S. Pappert (eds.), *Methods in empirical prosody research*. Berlin: Walter de Gruyter. 281–209.

Gibbon, D. 2009. "Can there be standards for spontaneous speech? Towards an ontology for speech resource exploitation". In: Tseng, S.-C. (ed.), *Linguistic Patterns in Spontaneous Speech*. (Language and Linguistics Monograph Series A25.) Taipei: Institute of Linguistics, Academia Sinica. 1–26.

Gibbon, D. and H. Richter (eds.). 1984. *Intonation, accent and rhythm. Studies in discourse phonology*. Berlin: de Gruyter.

Gibbon, D., I. Mertins and R. Moore (eds.). 2000. *Handbook of multimodal and spoken dialogue systems: Resources, terminology and product evaluation.* Dordrecht: Kluwer Academic Publishers.

Gibbon, D., U. Gut, B. Hell, K. Looks, A. Thies, and T. Trippel. 2003. "A computational model of arm gestures in conversation". *Proceedings of Eurospeech 2003*, Geneva.

Gibbon, D. and M. Kul. 2006. "Economy strategies in English and Polish text messages as examples of channel constraints". In: Jørgensen, J.N. (ed.), *Vallah Gurkensalat 4U & Me! Current perspectives in the study of youth language*. Bern: Peter Lang. 75–98.

Givón, T. 2002. *Bio-linguistics: The Santa Barbara lectures*. Amsterdam: Benjamins.

Grice, P. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Hall, E.T. 1959. *The silent language*. Garden City, NY: Doubleday.

Hauser, M.D., N. Chomsky and W.T. Fitch. 2002. "The faculty of language: What is it, who has it, and how did it evolve?" *Science* 298. 1569–1579.

Jakobson, R. 1960. "Closing statement: Linguistics and poetics". In: Sebeok, T. (ed.), *Style in language.* Cambridge, MA: MIT Press. 350–377.

Kay, M. 1987. "Nonconcatenative finite-state morphology". *Proceedings of the Third European ACL Conference*.

Kendon, A. 1972. "Some relationships between body motion and speech. An analysis of an example". In: Wolfe, A. and B. Pope (eds.), *Studies in dyadic communication*. Pergamon Press, New York. 177–210.

Kendon, A. 1996. "An agenda for gesture studies". *The Semiotic Review of Books* 7(3). 7–12.

Kendon, A. 2004. *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.

Kohler, K. 1987. "The linguistic functions of F0 peaks". *Proceedings of the 11th International Congress of Phonetic Sciences*, Tallinn (vol. 3). 149–152.

Lakoff, G. and M. Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Lashley, K.S. 1951. "The problem of serial order in behavior". In: Jefress, L.A. (ed.), *Cerebral mechanisms in behavior*. New York: John Wiley & Sons. 112–136.

Leech, G.N. 1983. *Principles of pragmatics*. London: Longman.

Lenneberg, E.H. 1967. *Biological foundations of language*. New York: John Wiley & Sons.

Levelt, W.J.M.1989. *Speaking. From intention to articulation*. Cambridge, MA: MIT Press.

Malinowski, B.K. 1923. "The problem of meaning in primitive languages". In: Ogden, C.K. and I.A. Richards (eds.), *The meaning of meaning.* London: Routledge. 146–152.

Martell, C. 2001. "Form: An extensible, kinematically-based gesture annotation scheme". *Proceedings of Language Resources and Evaluation Conference Proceedings (LREC 2001)*. 183–187.

McCullough, K.-E. 2005. Using gestures during speech: Self-generating indexical fields. (Unpublished PhD dissertation, The University of Chicago.)

McNeill, D. 1992. *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

McNeill, D. 2005. *Gesture and thought*. Chicago: University Of Chicago Press.

McNeill, D., F. Quek, K.-E. McCullough, S. Duncan, N. Furuyama, R. Bryll, X.-F. Ma and R. Ansari. 2001. "Catchments, prosody and discourse". *Gesture* 1(1). 9–33.

Nabokov, V. 1957. *Pnin*. Garden City, NY: Doubleday & Company.

Opie, I. and P. Opie and M. Warner. 1959. *The lore and language of schoolchildren*. Oxford: Oxford University Press.
Peirce, C.S. 1958–1960. *Collected papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.
Pittenger, R.E., C.F. Hockett and J.J. Danehy. 1960. *The first five minutes*. New York: Martineau.
Rossini, N. 2004. "Sociolinguistics in gesture: How about the Mano a Borsa?" *Intercultural Communication Studies*, XIII(3). *Proceedings of the 9th International Conference on Cross-Cultural Communication (CSF) 2003*. 144–154.
Rossini, N. In press. *Language in action. Reinterpreting gesture as language*. Amsterdam: IOS Press.
Rossini, N. and D. Gibbon. 2011. "Why gesture without speech but not talk without gesture?" *Proceedings of the International Congress of Phonetic Sciences (ICPhS) 2011*, Hong Kong.
Searle, J. 1969. *Speech acts*. Cambridge: Cambridge University Press.
Slobin, D.I. 2004. "From ontogenesis to phylogenesis: What can child language tell us about language evolution"? In: Langer, J., S.T. Parker and C. Milbrath (eds.), *Biology and knowledge revisited: From neurogenesis to psychogenesis*. Mahwah, NJ: Lawrence Erlbaum Associates. 255–285.
Thies, A. 2003. *First the hand, then the word: On gestural displacement in non-native English speech*. Bielefeld: SII thesis, Universität Bielefeld.
Trippel, T., A. Thies, K. Looks, U. Gut, J.-T. Milde, B. Hell and D. Gibbon. 1984. "CoGesT: A formal transcription system for conversational gesture". *Proceedings of the Language Resources and Evaluation Conference (LREC) 2004*.
Wachsmuth, I. and S. Kopp. 2002. "Lifelike gesture synthesis and timing for conversational agents". In: Wachsmuth, I. and T. Sowa (eds.), *GW 2001, LNAI 2298*. Berlin: Springer Verlag. 120–133.

Address correspondence to:
Dafydd Gibbon
Universität Bielefeld
Postfach 100131
33501 Bielefeld
Germany
gibbon@uni-bielefeld.de