Empirical Issues in Deictic Gestures: Referring to Objects in Simple Identification Tasks

Peter Kühnlein & Jens Stegmann

November 14, 2003

Contents

Abstract

The setting used for empirical studies pertaining to co-verbal pointing gestures is described. We outline steps in the annotation process and tools we employed. The observations we made in (i) the temporal and (ii) the spatial area are discussed. Some conclusions are drawn at the end of the report and some future prospects layed out.

Introduction

In this report we will describe some of the empirical studies we conducted in CRC SFB 360 B3 "Deikon" and discuss some of the results. The report is intended as the first of a series of reports, the second of which will explain details connected with the annotation, the third being theoretically oriented. What we will do here is lay out the empirical foundations of the theoretical work we will describe and discuss in the later report. A survey of the research is presented in (Kühnlein et al., 2003), a paper presented at the Gesture Conference at Austin, Texas (http://www.utexas.edu/coc/cms/International_House_of_Gestures/).

In chapter 1 we substantiate the claims on the connection between the parts of our work as they will be presented in this and the subsequent reports. This will be followed by a detailed description of the studies and their documentation in Chapter 2. We will describe the work on annotation we did and discuss alternatives to some extent in Chapter 3. Finally, we report findings we made during the studies that are related to our theoretical interests in Chapter 4.

We would like to stress the fact that our studies are by no means experimental in the strict sense of the word. Although we fixed some of the variables in our setting, and thus used quasi-experimental methods, the studies reported upon here have more the character of pre-tests. Recently, we have conducted a first follow-up of these studies. A more long-term agenda is to use virtual reality and eye tracking devices in addition to the traditional methods we employed in this study and will continue to use in the next. As Deikon is an interdisciplinary project comprising philosophers, linguists and computer scientists, there will be an opportunity to use some high-tech apparatus to test the philosophical theories we develop on the board and up to now checked in the empiricist's armchair.

Chapter 1

Objectives of the Studies

B3 "DEIKON" is set up to investigate phenomena of reference by deixis during construction dialogs[1]. The goal of DEIKON is to enable the Situated Artificial Communicator that is the objective of the CRC SFB 360 to deal with pointing in a (possibly virtual) environment during instruction phases. In the case of DEIKON, the Situated Artificial Communicator is realized as an embodied communicative agent that serves as a man-machine interface.

In order to enable the integration of speech and gesture, and having a symbolic AI background, DEIKON has started to develop a formal representation of the information carried by pointing gestures that is treatable by some standard symbolic linguistic mechanism. To detail the mechanism will be the purpose of another report. But some of the gross characteristics can be given here.

The mechanism has to be capable of rendering (and help explaining) the various temporal behaviors of gesture that have been observed in the literature and in our own studies. To render it, it has to be able to assign different places to the representation of the gesture in the representation of the discourse. To help explain it, it has to be able to describe the semantic effects of the various positions adequately.

Most prominent among the semantic effects are the failures of deictic pointing, the cases where the functions of the gestures are not fulfilled. These are most widely discussed in the semantic and philosophical literature on deictic gestures and demonstratives. But there are subtler effects that we will describe here and explain in the third of our reports.

Describing the temporal placement of the gesture formally consists, at least partly, in developing a syntactic format that allows for indicating symbolically the occurrence of the gestures. We will have such a representation in the follow-up paper that uses the south-east arrow ("↘") as a symbol for the stroke placement, cf. Rieser (2001). The representation will look something like X ↘ Y, and will be understood as "The gesture had its stroke between the surface portions[2] X and Y".

Of course, one has to know which are the temporal phenomena involved before it makes sense to develop a description or model. Most work on gestures that has been accomplished deals with a different kind of gestures, e.g., iconic or mimetic ones, and is not performed

[1] The acronym "DEIKON" stems from the German project title "**Dei**xis in **Kon**struktionsdialogen" which announces right that.

[2] The expression "surface portion" means that X and Y do not denote projections; rather they mean segments of the surface expression, independently of their grammatical status. We presuppose that the reader is acquainted with the standard partitioning of gestures as proposed or implied by various authors, e.g., Schegloff (1984) and Kendon (1980).

against the background of construction dialogs. The paradigm cases for studies of iconic gestures are narrative dialogs, e.g., the infamous "Tweety-and-Sylvester" story experiments, conducted by McNeill and his various co-workers[3] and reported in, e.g., McNeill (1992). But it seems reasonable to defend, or at least it is not *a priori* justified to reject, the assumption that pointing gestures behave differently from iconic gestures as time is concerned. This is due to the prima facie different function of deictic and iconic gestures. So, this is one of the goals of our studies: prepare the ground for the development of an adequate semantic model of speech plus deictic gestures in construction dialogs.

This ground-laying work has to do with another part of the description of the data: as the description in fact generates a corpus, it is desirable to have an annotation of the data that is by-and-large standardized. Standardization, in turn, is desirable for reasons of compatibility and re-usability of the corpus. As much time and effort was spent with finding a practical way of and format for annotation, during which many technical problems had to be mastered, we will describe this part as well, cf. Chapter 3. The third of our papers will show how the annotation is related to the grammar we decided to set up. It will be necessary to explain this connection, as XML, the annotational format we use, does not *per se* favor certain grammars. So additional restrictions have to be put on the XML annotation to express the grammatical character of the description.

Besides replicating the data from other experiments in our setting, the studies had a more ambitious aim. We hypothesized from the beginning that spatial constellations had an influence on the success conditions of the deictic gestures. Indeed we have strong indications that this is the case, suggesting that the semantic models which a complete theory has to employ need to respect more than just sets of objects. For modeling gestures in tandem with speech it seems that structured models have to be employed. Once again, of course, this will be one of the tasks that have to be achieved in the theoretical report that will be the follow-up of this one.

To sum up, we are reporting our results from the studies here, always having in mind that they are both exploratory and falsifying: exploratory in that they give us hints where we have to have a closer look at the normal behaviors and falsifying in that they tell us where to-date theories go wrong.

[3] Another prominent series of studies was conducted by Kita (1997); Kita et al. (1998), to name just some of the people working in the "Tweety-and-Sylvester" paradigm.

Chapter 2

Introducing the Setting

In this section, we describe the empirical setting that was used to establish the hypotheses. First, in Subsection 2.1, we state the traits of the setting and explain the recording. Then we describe how this setting differs from the standard setting that is used in the SFB 360, why we chose a non-standard setting, and what the consequences of the special features are.

2.1 Setting and Recording

In the empirical investigation that underlies our observations, we use a simplified version of the standard experimental setting of the SFB 360, as will be described in Section 2.2. Just like in the standard scenario, we have two subjects participating in the studies. One subject has a role we call *instructor* for compatibility during the investigation[1]. The duty of the instructor is to give the other subject, the *constructor*, enough information to identify one of the objects that lies on a table between them. The instructor can give the information, both verbally or gesturally or by some combination thereof. Indeed, instructor could even use other means of passing information if s/he liked. The constructor then takes the object and shortly lifts it from the table in order to show that s/he has identified it correctly. The objects that lie on the table are Baufix-parts that can be used to build a model airplane. Mid of the investigation subjects switch roles once.

The whole procedure is video-taped in a technically complicated way: we use two digital video cameras that are connected to a video mixer. The composite signal is then fed directly into a PC that is equipped with a special digital video card. This technique allows to store the video-streams in a completely synchronized way. We can view and handle the recordings from two (and in future, three) perspectives without having any delay between them or having to synchronize the videos after having recorded them. Besides this, we have the opportunity to simultaneously record the perspectives individually by using the built-in recording devices of the video cameras. And in addition, we can later electronically separate the composite signals if we need to have the single perspectives for annotation. The latter is interesting, because the frame-wise editing of the videos allows having starts and ends of the separated videos at exactly the same places, and hence they can be fed into a common time-base. Figure 2.1 shows a sample still from one of our studies, called “Pre-test 5”.

[1] In the standard scenario, the task of the instructor is to enable the constructor to build a toy airplane.

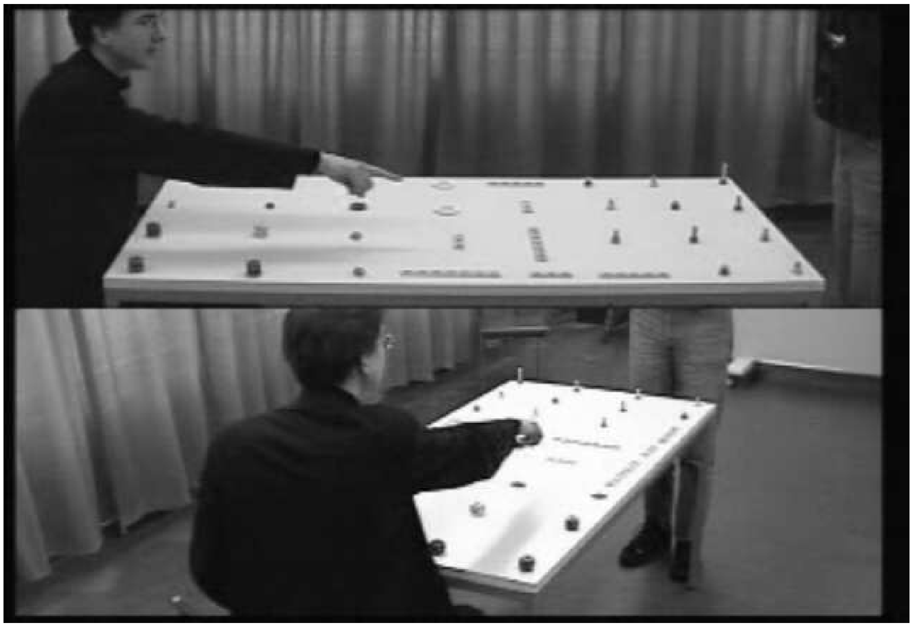


Figure 2.1: A sample still from pre-test 5.

The two perspectives of recording are that of instructor and one that covers the whole scene. Having both perspectives at our avail, we can easily reconstruct what the instructor saw, and how his/her movements developed. In future experiments, we will make use of three perspectives, covering also the view from constructor's side. Some of the pros and cons of the techniques used will be mentioned in Section 3.

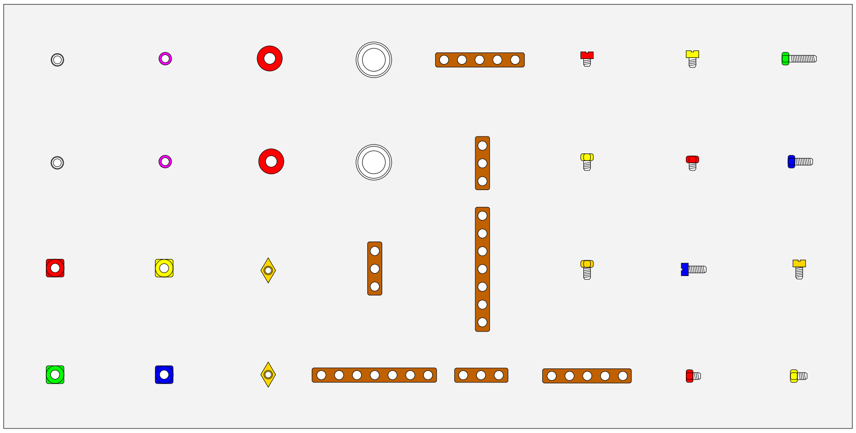


Figure 2.2: One of the two clusterings we used for the setting in top view: objects with similar shape are grouped together. In our studies, instructor and constructor were located opposite to each other at the shorter sides of the table. The instructor sat on a chair (simply to make videotaping easier), while the constructor could move in order to reach the scattered objects.

There are two clusterings of objects we use in order to avoid biases that could arise because of the layout of the objects: in the first clustering, objects with corresponding colors neighbor each other, in the other the criterion is sameness of form. The two clusterings are, however, not completely independent, as there are types of objects, e.g., the wooden bars, that share neither color nor form with any other type. Hence, objects of those types are bound to be clustered under either condition[2]. Figure 2.2 shows one of the clusterings we used for our studies.

We conducted seven studies with different subjects, i.e. we had 7 pairs of subjects (Ss) as a total. There were 32 objects on the table, and we had one change of roles per clustering method. Two clustering methods per study were used. This makes a total of $7 \times 32 \times 2 \times 2$ (= 896) identification tasks.

2.2 Special Features of this Setting

Our setting differs from the standard setting in the following respects: as the subjects need not cooperate on building a toy airplane, but merely on identifying atomic parts, no aggregation steps are involved. As a result, one of the crucial features of the SFB's basic scenario does not hold for the setting we chose for the studies. This both simplifies the task and allows some observations that could not be made otherwise.

First of all, there is no dynamic setting, i.e., the spatial distribution of the objects on the table does not change very much. The only spatial differences over time are the slight deviations from the original positions that occur when the constructor lifts the identified objects. These deviations are negligible, as will become clear below, for the present purposes. On the other hand, in the standard scenario massive changes of the location of the objects is part and parcel of the task. Spatial aggregation necessarily involves a dislocation of objects.

Secondly, as no aggregation steps occur, there is no necessity for the subjects to follow some plan for building the toy airplane. They can freely and randomly choose the objects that should be identified independently from the ones that were previously chosen. By contrast, the SFB's standard scenario is much more constrained as the sequential use of relevant objects is concerned. The results we gain from our studies are complementary to those that are specific for searching tasks.

A third reason to use a static setting with just pointing is that we can arrange the objects on the table according to different clusterings, e.g., color, form, or chance clustering. This allows to control possible preferences of choice according to features of the objects vs. their spatial distribution. The variation of clusterings is one step towards strict experimental conditions, as it amounts to fixing one of the conditions.

To sum up, the scenario differs in relevant characteristics from the standard scenario that is used in the CRC SFB 360; it proved that we were right in using this scenario for a start. Now that in doing so we narrowed down the search space for effects that can be expected, we can return to the standard scenario in follow-up studies. That means that aggregation steps and dynamic changes in the domain will have to be accounted for in the theory.

[2] In future settings we will therefore add a neutral, chance, arrangement of objects to the conditions.

Chapter 3

Annotation of Multi-Modal Phenomena among Gesture and Speech

> Linguistic annotation covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time-functions—audio, video, and/or physiological recordings—or it may be textual.
>
> (Bird and Liberman, 2001)

For our project, the following decisions were made with regard to annotation of the video data.

1. A *data storage and exchange format* had to be chosen. It had to provide means for the analysis and presentation of annotated data.

2. We had to think about applying an existing *annotation scheme* to our data. Among the possible alternatives are the adaptation of an existing proposal or, more radically, the formulation of a specification from scratch. Anyway, the chosen specification has to be translated (and, hence, be translatable) into the language of the data format (see 1. above).

3. An adequate *software tool* that helps facilitate the annotation of digital video data in real-time had to be applied. This tool had to support the data format (1. above) and allow for the realization of the annotation scheme (2. above).

The current section on annotation will be organized with respect to these three topics. It should be stressed that the indicated order (1.–3. above) is somewhat arbitrary and does not have to be followed in practice. For example, we might have made decisions in favor of a certain software-tool (3.) first, whereby the range of options for annotation schemes (2.), as well as for data formats (1.) would have been accordingly constrained. Our elaborating remarks above are to be considered from an appropriate perspective.

In our presentation, we will deal with the issues of annotation schemes (2.) and software tools (3.) in Section 3.2 simultaneously. This happens in order to reflect “historical facts” concerning our orientation toward the respective points.

3.1 XML as Data Format for Multi-Modal Annotation

In fact, our first decision was in favor of the (meta-)markup language XML (*eXtensible Markup Language*) as the data storage and exchange format of our choice. We will elaborate on our reasons for this decision in Section 3.1.1. This discussion will be followed by remarks on special problems that arise with the application of XML in the domain of digital video data and phenomena of multi-modal communication in Section 3.1.2.

3.1.1 Reasons for XML-based Annotation

The XML standard is at the heart of a rich software-technological infrastructure encompassing a wealth of processing software and supplementing technologies that build upon it. Among those are XML applications in a narrow sense, i.e. applications that are themselves realized in a generic XML syntax. Most importantly, with respect to our purposes, XSLT (the *eXtensible Stylesheet Transformation*) marks the application of XML to define a functional programming language that is specifically designed to bring about the transformation of XML documents. Generally, XML is a uniform framework for a broad range of applications, with a coverage from narrative-centric to data-centric document formats.

It should be noted, however, that XML is mainly geared toward the processing of textual data. In the first line, it provides means for the use of markup-like syntax which can be tailored to meet the demands at hand.[1] Compliant markup is realized *via* insertion of start and end tags of elements at appropriate positions in a document instance. Furthermore, element start tags may get augmented with attribute value pairs in order to state (meta-)information, i.e. information about the respective element tokens at hand. Elements and attributes are among the basic inventory for primary information structuring with XML. One of the most distinctive features of XML is its underlying data format: the tree structure. This means that the information units in a well-formed document must, in honor of obedience to the generic syntax of XML, be representable as a tree. This feature is an advantageous property in many domains. Nevertheless, it is also responsible for certain problems associated with the use of XML in yet other domains, cf. 3.1.2.

Documents that satisfy the generic XML syntax are said to be *well-formed* documents or document instances. Markup can be further constrained by delivering a document grammar in form of a DTD (*Document Type Declaration*). An XML document that respects all the constraints laid down in a DTD is called *valid* with respect to that DTD. Validity, unlike well-formedness, is optional within XML[2].

We decided that this would not be a good place to go into the nifty details of XML and supplementing technologies. Instead we will concentrate on the most important points with respect to the use of XML for means of multi-modal annotation in the domain of digital video data. To do this, we have to presuppose some level of acquaintance with XML (and possibly

[1] Superficially, XML markup bears a close resemblance to HTML markup, except that the tag names usually are different. It should be noted that standard HTML builds upon XML's predecessor technology, SGML (*Standardized General Markup Language*). By providing meta-language functionality both XML and SGML allow for the specification of object markup languages such as HTML. XML can be conceived of as a subset of SGML—that subset which results by dropping all more or less exotic features that stand in the way of efficient implementation without touching general expressivity. There is also an HTML-version that builds upon XML (X-HTML).

[2] This holds in marked contrast with SGML, XML's predecessor technology.

XSLT). For further reading, please consult the pertinent literature (hints are given at the end of this section).

Some people concerned with annotation argue in favor of formats that are compliant with relational data bases instead of XML.[3] But with the availability of XML-based tools for annotation, XML-based technologies for further processing—especially XSLT designed for further transformation—and, perhaps most importantly, XML itself becoming the *lingua franca* within the research community, we refrained from taking that option. Anyway, tempers have run down a bit meanwhile. This has to do with the fact that the technologies have shown to be complementary in certain respects. Today's state-of-the-art systems tend to be hybrid in that they try to marry the respective advantages of both approaches: *(i)* "internal" storage in a relational database format for the sake of efficiency (load/save, search, query), and *(ii)* XML as "external" exchange and interchange format (*interlingua*, readability), as well as for further processing means to saddle upon (e.g. transformation for means of analysis and presentation). Examples are TASX-ANNOTATOR (Milde and Gut, 2001), to be covered in Section 3.2.4, new developments underway within the ATLAS project (Bird and Liberman, 2001), and the NITE workbench (Bernsen et al., 2002).

3.1.2 Problems with Digital Video Data and Multi-Modality

There are special problems with the employment of XML as an annotation format for non-textual information (e.g. our digital video data). With textual data, the content of an element usually comprises a sequence of (parsed) character data, i.e. text. Analogously, with digital video data one might think of short intervals of video to show up between the respective start and end tags. An approach along such lines might get implemented by first declaring appropriate *notations* for the various data types to be included (e.g. AVI). The digital video data would have to be encapsulated in external files along with the declaration of respective entities. Finally, presumably empty elements that would have to go with attributes taking appropriately constrained values would have to be employed. These would figure as entities for means of reference to the external data files. At the level of the document instance, those elements would appear as the content of classifying elements. They would thereby fulfill their role as place-holders for the original digital video data.

Note, first, that the digital video data would still go by the status of unparsed external entities (compared to the status of parsed character data in the motivating treatment of textual data). Over and above that, such an approach would still not do, at least not in a straightforward way. The reason is that it would be most arbitrary to choose any particular duration for the episodes of digital video, apart from one option: to ensure sustainability, one would have to specify an external data file (as well as a corresponding entity) for each and every single frame that the corpus comprises. If one chose otherwise, in case the need for more fine-grained annotation arose during the course of research, it would be necessary to re-do the implementation steps described above every time the granularity of time-based analysis changes. A procedure like this would be highly uneconomic. The prospects of taking such a *file-per-frame approach* seemed doubious to us—especially with an eye toward maintainability and further processing means—and, for good reasons we suspect, we are not aware of anyone pursuing this kind of strategy.

[3] Being in danger of gross oversimplification: the main opponents in the respective 'debate' were the American ATLAS project (using a relational data base compliant format) and the European MATE research consortium (favoring XML-centered solutions).

But there is a more general problem pertaining to information modeling with XML lurking behind—one that is related to XML's above mentioned tree structure format. It is an ubiquitous phenomenon that entities belonging to different modalities (e.g., words and gesture phases) overlap in all conceivable ways. They are rarely, if ever, occurring strictly parallel in time. Naïve XML annotation of such phenomena leads to serious problems with the underlying tree format, since a tree, by definition, does not allow for intersecting elements.

However, there is a standard solution with XML: the use of a strategy commonly known as *stand-off annotation* ((Sperberg-McQueen and Burnard, 1994), (Thompson and McKelvie, 1997)). In stand-off annotation elements containing attribute-value pairs are used to express linkages and bring about the desired function. Only those elements that are referred to by such links contain the desired data as their literal element content. By contrast, the elements that contain the referring link will be empty.

In a basic version, stand-off annotation thus can be realized *via* usual XML linking mechanisms that operate on the level of the document instance. "Usual linking mechanism" means "pairs" of specifically constrained attributes, taking respective ID and IDREF values.

A second, more powerful variant of stand-off annotation makes use of supplementing XML technology, notably *XPointer* and *XLink*. This technique allows expressing more complex linkages, e.g., to a whole range of XML information units. The related units may even be placed outside the current document instance.

What might be considered as realizing a third variant of stand-off annotation consists in the specification of exact timing information by the attachment of time-stamp attributes[4]. The difference with respect to the first and the second variant just described is as follows: the latter uses values that refer to information units which, in turn, contain the desired data as their element content. The third approach, however, makes use of a numerical value that corresponds to a point of time (or a picture frame). It is evident that the points of time (or frames) can be related to the foundational video data.

Since the relevant information can be encoded with stand-off annotation, the problem of overlapping elements is resolved. Of course, one needs a special processing software to bring about either approach. The annotation tools that will be described in Sections 3.2.2 and 3.2.4 below are employing stand-off annotation in the sense of the last variant described, whereas the *Mate Workbench* mentioned in Section 3.2.1 makes use of the second approach.

3.1.3 Relevant Literature

To close this section on XML for multi-modal annotation, we list some of the relevant literature.

For an introduction to XML one may consult (Ray, 2001). (Lobin, 2000) is also introductory, but focuses on primary and secondary information-structuring with XML and SGML. (Harold and Means, 2001) functions primarily as a reference, but note that there are introductory chapters that cover the core language standard, as well as supplementing technologies, most importantly XSLT and XPATH.

The general strategy of using XML together with XSLT for linguistic annotation is elaborated upon in (Carletta et al., 2002).

The most important alternative approach is the *annotation graphs* framework (Bird and Liberman, 1999), especially the developments underway within the ATLAS project (Bird et al.,

[4] One for the start and one for the end time in order to indicate that interval which comprises the phenomenon at hand.

2000). Although developers working within this paradigm show affinity toward relational data-base compliant formats, they also go with an XML-based interchange format.

Historically, in opposition to the ATLAS perspective, the MATE project, e.g., (Dybkjær, 2000), has been the most influential advocate of XML-based annotation.

3.2 Schemes and Tools for Annotation

In this section, we will introduce three possible annotation schemes and appropriate software tools that can be used in combination with XML. We introduce them in the order we experimented with them. The first two, to be dealt with in Sections 3.2.1 and 3.2.2, can be classified as top-down. Together with these, DTDs, or specification files, play a prominent role in constraining annotation in relevant ways from a theory-driven perspective. The introduction of these tools will be followed by an elaboration on the respective advantages and disadvantages of top-down and bottom-up strategies regarding scientific annotation. The stance we finally adopted in our project is essentially bottom-up and will be presented in Section 3.2.4.

3.2.1 Approach I: Top-Down Style with MATE

At the beginning of our project, there were no appropriate XML-based standardization proposals with regard to the annotation of multi-modal signs (e.g., gesture plus speech). We decided to design and implement our own annotation scheme.

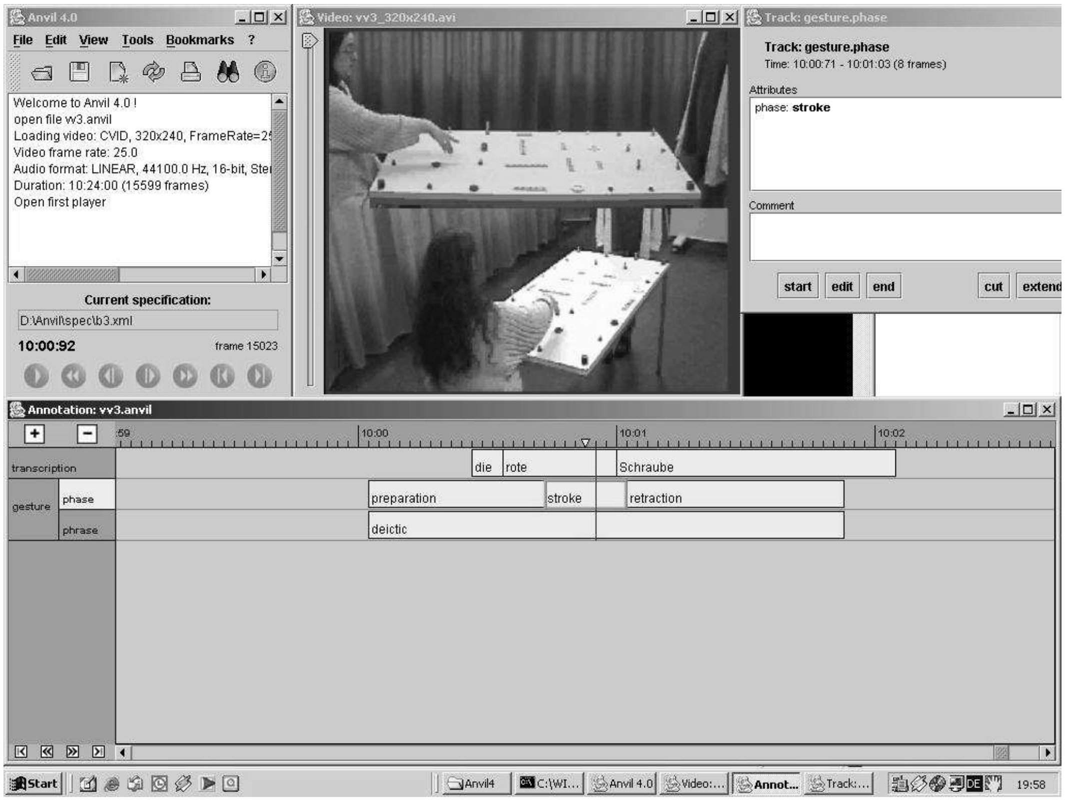
We wanted our specification to be compatible with the framework that was under development within the MATE project. This seemed to be the most promising approach toward a general architecture for multi-modal annotation then. Our plan was to come up with coding modules, i.e. well-documented, modular DTDs, in the sense of the respective MATE meta-specifications (Dybkjær et al., 1998). We started writing several DTDs: one for gesture form, basically a complete implementation of the gesture description framework laid out in (McNeill, 1992), and one for the annotation of dialog acts following the approach outlined in (Carletta et al., 1996). This approach, however, proved to be unviable for two main reasons.

Foremost, we realized that we would need prophetical foresight with respect to the phenomena that we might wish to annotate later on during our course of research—compare the critique of the top-down, theory-driven approach toward annotation below in Section 3.2.3. Furthermore, we ended up without an adequate software tool to facilitate the MATE-style of multi-level annotation. The MATE workbench, see (McKelvie et al., 2001), proved to be unstable and, furthermore, provided no facilities for real-time based annotation of digital video data.

3.2.2 Approach II: Top-Down Style with ANVIL

ANVIL (homepage: `http://www.dfki.de/~kipp/anvil`) is a free annotation tool that has been designed for the processing of digital video data (Quicktime or AVI files). It is written in *Java2* with *JMF 2.1.1* and can therefore be used on any platform that is supported by the latter. It is easy to install and provides a comfortable working environment in the form of a graphical user interface that allows for different layers (temporal, hierarchical and user-defined). Furthermore, it offers project management functionalities, as well as export to statistical processing software, e.g., *SPSS*, and import from selected tools (*Praat*, *XWaves* and *RSTtool*).

ANVIL's underlying annotation framework is object-oriented and allows for a broad range of annotation schemes. An annotation scheme that shall be employed must be licensed *via* implementation in the form of a *specification file*. A specification file is, in effect, an XML document that functions as an "annotation grammar" in the ANVIL context, comparable to the role of DTDs in the general XML context. ANVIL annotations of video data are guaranteed to be valid with regard to their specification files, since respective parameters get appropriately constrained. The use of a specification file is obligatory and it is exactly this feature that is responsible for ANVIL's marked top-down character. Finally, annotation results get stored in the form of an ANVIL *data file* which is, again, an XML file.



```
Welcome to Anvil 4.0 !
open file vv3.anvil
Loading video: CVID, 320x240, FrameRate=25
Video frame rate: 25.0
Audio format: LINEAR, 44100.0 Hz, 16-bit, Ste
Duration: 10:24:00 (15599 frames)
Open first player
```

Current specification: D:\Anvil\spec\b3.xml

10:00:92 frame 15023

Track: gesture.phase

Time: 10:00:71 - 10:01:03 (8 frames)

Attributes

phase: **stroke**

Comment

Figure 3.1: Basic annotation with ANVIL

Figure 3.1 shows a screen-shot from our ANVIL-annotated data. The annotation concerns the variety of temporal/structural relationships among gesture and speech that could reasonably be expected in multi-modal utterances in our task-oriented setting. It proved to be sufficient to apply a rather simple annotation scheme with the following three layers: gesture.phase, gesture.phrase and transcription. Again, the gestural layers are informed by McNeill's description framework (McNeill, 1992). The relevant categories will be treated in detail in Section 4.1[5]. Note that no information concerning gesture morphology (hand posture, finger orientation etc.) has been included. This was done deliberately, in order to keep the annotations as simple as possible. The results to be reported in Section 4.1 are based on

[5] The XML annotation data file comprising the extract displayed in Figure 3.1, as well as the respective XML specification file are available for download.

data annotated in this way.

ANVIL is an excellent tool for the pursuit of a top-down approach to annotation. However, such a strategy comes at certain expenses, as will be discussed in the following section.

3.2.3 Discussion: Top-Down vs. Bottom-Up Approaches

The approaches laid out in the preceding sections essentially follow what one might call a *top-down* or *theory-driven approach* toward annotation: the phenomena to be annotated are specified first from a theoretical perspective, thereby constraining the range of relevant phenomena to be annotated. Only at this point processing means are tailored to fit that specification.

Such an approach has the advantage of delivering clear-cut directives and theoretical orientation for annotators. This is achieved by specifying which phenomena are relevant and how they are to be encoded in the annotation file. In other words, one can specify an annotation strategy (cf. Mitkov (2002)) to a high degree. In fact, coding books (or similar guidelines) which prescribe how and what to annotate are exactly manifestations of such top-down approaches. Another positive feature is that the syntax of the annotated data may be checked for validity, i.e. whether they are compliant with the underlying specification. This is most obvious with DTDs.

But, on the other hand, following this type of strategy means that modifications of the foundational design generally tend to be problematic. This is because every time the underlying specification gets changed—which might occur owing to shifts in theoretical perspective or focus—annotated data and processing software have to be changed accordingly. So, every modification results in error-prone maintenance work.

An alternative is a more *bottom-up* or rather *data-driven approach* toward annotation. To pursue such a strategy could mean two things in terms of XML: first, one could try to do without a DTD altogether, thereby gaining a high degree of flexibility with respect to annotation. The drawback of this decision would be a lack of structure in the annotated data: the only structure would stem from the bare XML skeleton. And this would be too little to build processing software (e.g., XSLT scripts) upon.

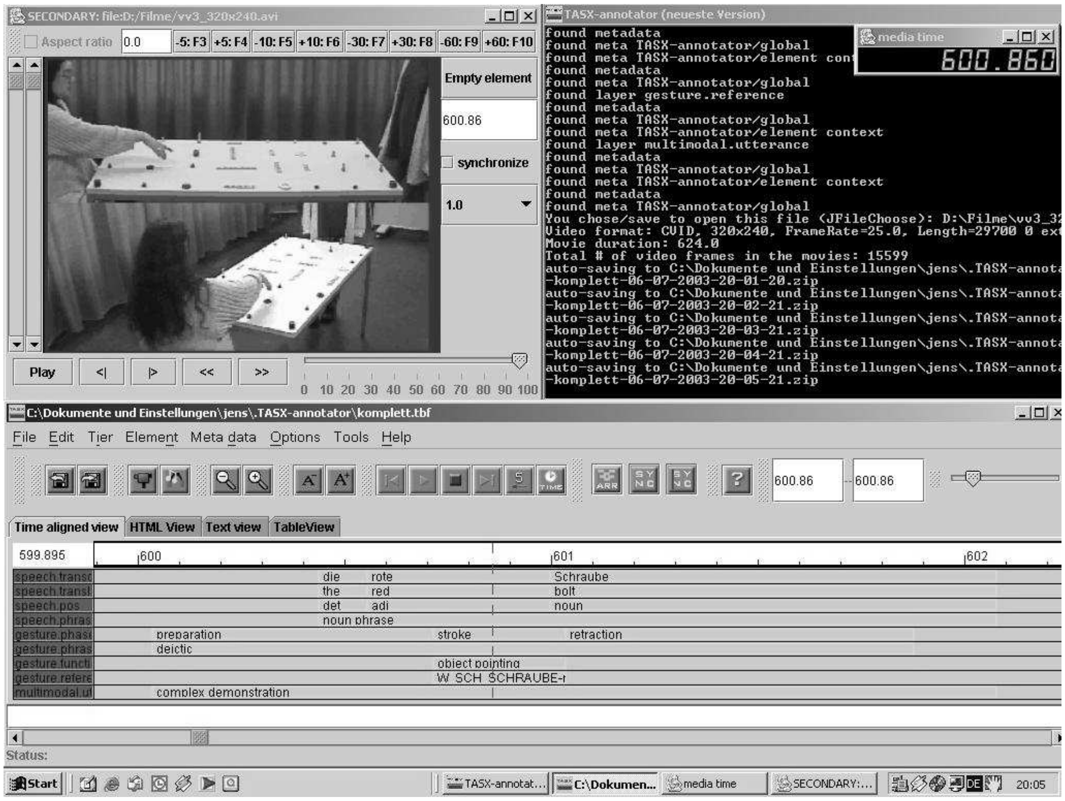
A second, more promising, option is to have a DTD to constrain the annotation of data, but to write it in a way that allows for a certain degree of flexibility. This goal can be attained by means of rather flat hierarchies and the employment of element content, rather than element type, to encode the information. Such an approach could count as a compromise between radical top-down and bottom-up approaches. In spite of its less rigorous character and the option to incorporate empirical insights in an easy way, we nevertheless subsume it under the heading of bottom-up approaches. Anyway, this approach seems to do more justice to the dialectic character of empirical research (with top-down and bottom-up perspectives usually being intertwined dynamically).

3.2.4 Approach III: Bottom-Up Style with TASX-ANNOTATOR

In order to overcome the limitations of the top-down approach to annotation, tools are needed that work without predefined schemata. The tool of our choice was the TASX-Annotator (homepage: `http://tasxforce.lili.uni-bielefeld.de`), which is a central component of the TASX-environment (Milde and Gut, 2001)[6]. It allows for the annotation of multi-channel

[6] TASX is short for the *Time Aligned Signal data eXchange format*.

digital video (and audio) data. Different views for annotated data are provided. TasX-Annotator is written in *Java2* with *JMF*, and, apart from this, makes use of xslt and *Perl* scripts. A set of transcoder tools realize import/export from a range of formats (including the Anvil data format). Annotations are realized in agreement with the TasX-dtd. This flexible format is a good example of realizing an xml-based bottom-up approach in the sense of the preceding section. However, internal storage is handled in a relational database for the sake of efficiency. Furthermore, it should be noted that the software—in contrast to Anvil—is open source and distributed for free under the auspices of the *GNU General Public License*.



```
found metadata
found meta TASX-annotator/global
found meta TASX-annotator/element con
found metadata
found meta TASX-annotator/global
found layer gesture.reference
found metadata
found meta TASX-annotator/global
found meta TASX-annotator/element context
found layer multimodal.utterance
found metadata
found meta TASX-annotator/global
found meta TASX-annotator/element context
found metadata
found meta TASX-annotator/global
You chose/save to open this file (JFileChoose): D:\Filme\vv3_32
Video format: CVID, 320x240, FrameRate=25.0, Length=29700 0 ext
Movie duration: 624.0
Total # of video frames in the movies: 15599
auto-saving to C:\Dokumente und Einstellungen\jens\.TASX-annota
-komplett-06-07-2003-20-01-20.zip
auto-saving to C:\Dokumente und Einstellungen\jens\.TASX-annota
-komplett-06-07-2003-20-02-21.zip
auto-saving to C:\Dokumente und Einstellungen\jens\.TASX-annota
-komplett-06-07-2003-20-03-21.zip
auto-saving to C:\Dokumente und Einstellungen\jens\.TASX-annota
-komplett-06-07-2003-20-04-21.zip
auto-saving to C:\Dokumente und Einstellungen\jens\.TASX-annota
-komplett-06-07-2003-20-05-21.zip
```

599.895	600				601	602
speech.transc		die	rote		Schraube	
speech.transl		the	red		bolt	
speech.pos		det	adi		noun	
speech.phras		noun phrase				
gesture.phase	preparation			stroke	retraction	
gesture.phras	deictic					
gesture.functi				obiect pointing		
gesture.refere				W SCH SCHRAUBE-r		
multimodal.ut	complex demonstration					

Figure 3.2: Augmented annotation with TasX-Annotator

With TasX-Annotator, new tiers can be invented (and existing ones re-ordered) at any time during annotation. For instance, we utilized this feature when we supplemented our basic annotations with additional layers of information.

In a first step, we added a *translation* tier (English translation), mainly for presentational purposes. In a second step, we augmented our annotated data more decisively in order to incorporate theoretical insights and to meet the demands of an external specification. The plan was to transform part of the annotated data from our empirical studies so that they could be used as (benchmark-)scripts for the generation of multi-modal communicative behavior in a virtual reality setting (Sowa and Wachsmuth, 2002). In this setting, the respective behavior has to be specified by means of the xml-based specification language Murml[7] (Kranstedt et al., 2002a). In the course of this enterprise, the addition of new informational tiers went

[7] Murml is shorthand for *Multimodal Utterance Representation Markup Language*

hand in hand with the development of an XSLT stylesheet for the transformation from our TASX annotation format to the MURML format.[8] Figure 3.2 shows a screenshot from the respectively augmented TASX-annotated data.

Future annotation efforts will continue to make use of the TASX-ANNOTATOR, which allows for our current annotation scheme to be expanded and modified, as needs arise. One application is the annotation of details of gesture form in order to achieve a more fine-grained synthesis in the virtual reality setting.

3.2.5 Relevant Literature

An excellent source on information is the "linguistic annotation" homepage, which can be found at `http://www.ldc.upenn.edu/annotation`. By now, there is also a "gesture annotation" homepage: `http://www.ldc.upenn.edu/annotation/gesture`.

Recently, some interesting proposals concerning the XML-based annotation of co-verbal gestures have been made. For example, (Wittenburg et al., 2002) summarizes the work undertaken at the MPI for Psycholinguistics in Nijmegen, especially the MPI movement phase coding scheme (Kita et al., 1998) which can be integrated with the MPI articulator coding scheme (Kita et al., 2000). As is the case with MURML (Kranstedt et al., 2002a), this work incorporates ideas from McNeill's (McNeill, 1992) and the HAMNOSYS (Prillwitz et al., 1989) framework. The kinematically-oriented FORM scheme (Martell, 2002) is rooted in the paradigm of annotation graphs.

Follow-up projects to MATE, namely NITE (*Natural Interactivity Tools Engineering*) and NIMM (*Natural Interaction and MultiModality*), have taken up the MATE-line of research with focus on multi-modal phenomena, especially gesture and facial expression. Thereby NIMM (Dybkjær and Bernsen, 2002) focuses mainly on documentation and evaluation of existing data resources, annotation schemes and tools, while the NITE project (Bernsen et al., 2002) aims at the development of a best-practice software tool for the annotation of digital video data.

[8] Relevant XML documents (annotation example, XSLT-stylesheets) are available for download.

Chapter 4

Observations

There were two classes of phenomena we concentrated on. The first class, dealt with in subsection 4.1, pertains to temporal restrictions that must be obeyed in multi-modal discourse. Much previous work has been done previously in this area and we will try to replicate some of the results. The other class has not been covered yet in empirical studies. What we could observe nicely in our studies were the spatial relations that have to obtain in order for pointing gestures to be selected as a means in communication and to be successful. These observations will be reported in Subsection 4.2.

4.1 Temporal Phenomena

In this chapter, we will report findings on temporal phenomena among gesture and speech. Our observations will be discussed against a background of psycho-linguistic observations and assumptions, specifically aspects of the framework formulated by McNeill (1992).[1] We will refer to his position with regard to timing, which integrates a bulk of research undertaken by McNeill and others (most importantly Kendon (1972, 1980) and Kita (1990)), as *the canonical view*. It is the dominating position in the field of gesture studies today[2].

However, besides from being of interest for psycholinguistic theories of gesture- and speech-production, there is another good reason for looking at temporal relationships. This one has to do with the main theoretical aim of our project, which is the development of a representational account of the impact of pointing gestures in multi-modal directives, as can be observed in task-oriented dialog (Kühnlein et al., 2003). Empirical findings on temporal succession have a profound influence on respective points, e.g. pertaining the range of admissible syntactic configurations.

The organization of this section will be as follows: First, we present a summary of pertinent aspects of the framework that has been set up by David McNeill in the next subsection. In

[1] Note that we do not intend to take a stance with respect to McNeill's underlying theory of gesture and speech production (*"growth point theory"*) nor the underlying neo-Vygotskyan research program (Hanfmann and Vakar, 1962) here.

[2] With respect to the space of theories and findings concerning gesture and speech, one may want to read an overview article, for example (Rimé and Schiaratura, 1991). Among the most important sources of research figures the newly-created journal "Gesture" (Kluwer Academic Press) and the online publications at the International House of Gestures (`http://www.utexas.edu/coc/cms/International_House_of_Gestures/`). A representative collection of papers is (McNeill, 2000).

doing so, we will introduce the necessary vocabulary and theoretical background against which we can discuss our own observations in a second step.

4.1.1 Background: McNeill's Framework

Unless indicated otherwise, the material in the following subsections is taken from (McNeill, 1992). We will begin our summary by reporting on McNeill's perspective with respect to hierarchical and temporal structure. This will be followed by remarks on the canonical view with regard to gesture anticipation and synchrony.

Hierarchical and Temporal Structure

McNeill describes a hierarchy of gestural movements as illustrated in Figure 4.1

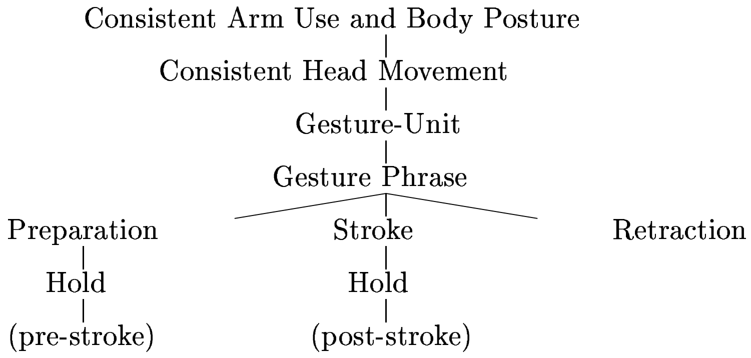


Figure 4.1: Hierarchy of gesture movements [from (McNeill, 1992, p.82)]

The uppermost hierarchical levels are *consistent arm use and body posture*, and, one level beneath, *consistent head movements.*[3] This means that one instance of 'consistent arm use and body posture' can comprise any number of 'consistent head movements'. A unit of the latter may contain several *gesture units*, each of which will span that period of time between successive rests of the limbs. Descending one more level of the hierarchy, several *gesture phrases* may cluster to form a gesture unit. Note that the concept underlying units on the level of the gesture phrase comes closest to the common-sense concept underlying the use of "gesture" in ordinary language. McNeill discriminates basically four, at times, however, no less than six different gesture phrase types: *iconics*, *metaphorics*, *deictics*, and *beats* (plus *cohesives* and *butterworths*).[4] We won't go into the details of gesture phrase classification, since we will be concerned exclusively with pointing gestures here, i.e., *deictics* in McNeill's terminology:

> "*Deictic* gestures are pointing movements, which are prototypically performed with the pointing finger, although any extensible object or body part can be used,

[3] Compare his respective remarks: "Within units on this level [body posture and arm usage], the speaker adopts different body postures and arm usage patterns. (...) Shifting between arm options and body postures defines a kinesic unit on this level. (...) Within stretches of a consistent arm and body use shorter stretches occur in which the same head movements take place." (McNeill, 1992, p.82)

[4] Compare the typologies in Chapters 1 and 3 of (McNeill, 1992).

including the head, nose, or chin, as well as manipulated artefacts." (McNeill, 1992, p.80)

According to Figure 4.1, gesture phrases do have *movement phases* as their parts. McNeill describes them roughly as follows:

Preparation: The limb moves away from its rest position to a position in gesture space where the stroke will take place.

Stroke: The peak of effort in the gesture. In this phase the meaning of the gesture is expressed. Strokes are typically performed in the central gesture space bounded roughly by the waist, shoulders, and arms. The head becomes involved occasionally.

Retraction: The return of the hand to a rest position, which does not have to be the one occupied before the current gesture phrase.

Hold: Any temporary cessation of movement without leaving the gesture phrase.

Table 4.1: Description of movement phases, (McNeill, 1992, p.83). The description is repeated here with slight modifications and abbreviations.

Gestures (apart from beats) are usually tripartite structures consisting of a preparation phase, followed by a stroke and, finally, a retraction phase. The gestural stroke is defined to be the meaningful, obligatory part of every gesture—all other phases are, at least in principle, optional. However, preparation phases are rarely, if ever, omitted. Retraction phases may be missing when a gesture passes directly into a successor gesture. Preparation phases and strokes may be followed by a hold (*pre-stroke* and *post-stroke holds*, respectively), whereby preparation and hold phases can be attuned in order to achieve synchrony with linguistic affiliates (more on this topic in the next section). The screenshot in Figure 4.2 depicts a prototypical deictic stroke from our data.

"The Canonical View": Anticipation and Synchrony

> "(...) gestures and speech have a constant relationship in time."
>
> (McNeill, 1992, p.25)

The above quotation summarizes the core idea of the canonical view. More specifically, it is postulated that gestures are to *anticipate* and *synchronize* with their linguistic affiliates. While the former means that gestures will usually go before coexpressive linguistic segments in time, the point of the latter is that gestures occur at the same time as their linguistic counterparts. This seems to be a paradox at first sight. The problem disappears, as soon as the tripartite structure of a prototypical gesture is taken into account. Anticipation and synchrony are meant to be related to different movement phase types: "The synchrony rules refer to the stroke phase: anticipation refers to the preparation phase" (McNeill, 1992, p.26).

Gesture Anticipation: If there is a preparation phase (...) [it] regularly anticipates by a brief interval the co-expressive linguistic segments. (McNeill, 1992, p.25)

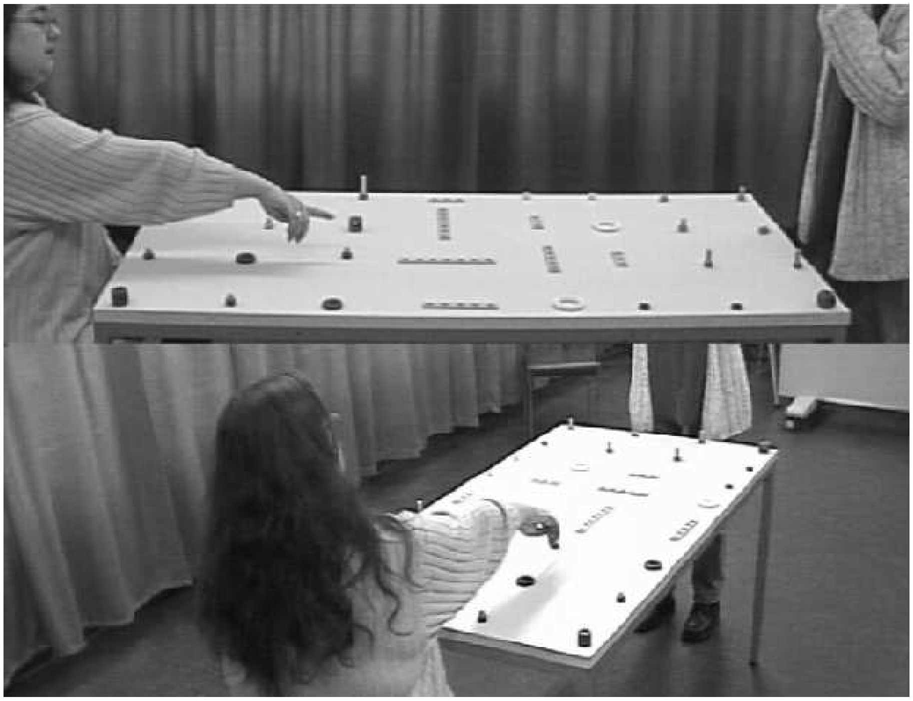


Figure 4.2: Deictic gesture stroke

Conceived of this way, anticipation can be seen to be a pre-requisite for gesture synchrony, as formulated below. Nevertheless, anticipation is stated as a flexible constraint (note the use of "regularly"). Therefore, McNeill's way of formulating his point seems to allow for exceptions from the rule.

Gestural synchrony, by contrast, is stated in the form of three law-like rules that do obviously not allow for exceptions. Each rule appeals to a specific level of linguistic analysis. Note, that there is no rule of *syntactic synchrony*, as might be expected:

Phonological Synchrony: "The synchrony rule at this level is that the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech (Kendon, 1980). In other words, the stroke phase of the gesture is integrated into the phonology of the utterance." (McNeill, 1992, p.26)

Semantic Synchrony: "Semantic synchrony means that the two channels, speech and gesture, present the same meanings at the same time. The rule can be stated as follows: if gestures and speech co-occur they must cover the same idea unit. The term 'idea unit′ is meant to make provision for synchronized speech and gestures where the meanings complement one another." (McNeill, 1992, p.27)

Pragmatic Synchrony: "The rule here says that if gestures and speech co-occur they perform the same pragmatic functions. Pragmatic synchrony implies that speakers are limited to one pragmatic reference at a time." (McNeill, 1992, p.29)

4.1.2 Own Contributions: Problems, Findings and Preliminaries

In this section, we report findings based on data from our first line of studies (the "pre-tests"). Some qualifying remarks are in order: the whole line comprised of eight pairs of subjects. Most of the subjects did not make use of pointing gestures trying to fulfill the task we set for them. If we had intended to do serious statistical evaluation, we would have been faced with a data basis that was too small. However, our intention was to use the data as "intuition pumps" to break the bounds of academically biased thinking (e.g., by the canonical view) about the phenomena at hand. Hence our presentation will take the form of a case-study with all the data coming from pre-test #3. Despite the limited range of data, we will not restrict ourselves solely to qualitative analysis, as might be found appropriate in the context of a case study. Therefore, where quantitative data are given, it has to be kept in mind that they stem from just one run of the study.

On the grounds of superficial analysis, pre-test #3 was judged containing the most interesting timing phenomena. The respective data serve to illustrate the diversity of timing phenomena that may reasonably be expected to appear in a task-oriented setting—something which our theoretical apparatus (Kühnlein et al., 2003) should be able to handle.

Problems that arose due to the application of McNeill's movement phase classification scheme will now be discussed. Right after that we will draw a detailed picture of our findings against the background of the canonical view. At the end of this section, we will state some preliminary thoughts concerning our need of bridging the gap from temporally extended empirical phenomena to a linear symbolic representation for our linguistic formalism to operate upon.

Problems in the Application of McNeill's Phase Typology

The first thing to note is that the claims made by the canonical view that pertain to optional ommission of preparation and retraction phases accord with our findings. However, we are not in a position to say anything substantial about holds. This has to do with a foundational problem underlying McNeill's typology of movement phases: it does not cleanly discriminate semantic and kinematic criteria. With deictic gestures, the differing perspectives lead to boundaries drawn in different ways. Figure 4.2 shows a prototypical deictic gesture stroke. At this stage, position, shape and posture of the gesticulating hand and arm remain relatively stable[5] for a moment (respective values are given on page 21). From a kinematic point of view, this "non-movement" phase qualifies as post-stroke hold. But, from a semantical perspective, during this phase the gesture expresses its meaning function and therefore classifies as stroke.

In the annotations underlying our findings, we followed the semantic strategy. Thereby, we made use of the following rule of thumb: the stroke begins as soon as the index finger "comes to rest" with respect to the region or objects pointed at—the stroke ends with the initiation of the retraction movement. The region/object pointed at then becomes indetermined. We are aware that this convention is rather vague, but we were unable to come up with a better one at the time of annotation. Furthermore, under a semantic interpretation of stroke, detection of stroke boundaries will probably always be a matter of discussion and should be decided on a case-wise basis.

We are aware of the somewhat paradoxical nature of a semantic approach to stroke classification, as "stroke" is terminology which seems to belong under the heading of gesture form

[5] Some minor correction movements may occur.

and should therefore be judged taking into account form measures as hand position, posture, velocity of movement and the like. Nevertheless, there has to be some sort of correspondence mapping between form-based categories and semantic units. Anyway, we plan to adapt to a kinematic perspective[6] in further investigations. Note, however, that a respective switch will not undermine the general force of our claims in the sections below.

Findings Related to the Canonical View

Gesture anticipation and *synchrony* were investigated with respect to XSLT-processed data from our time-line based XML annotation files. The algorithms and variables exploited for transforming our species of TASX to MURML (as mentioned at the end of section 3.2.4 and to be described in a sequel to this report) proved to be useful and flexible enough to be adapted and modified for means of analysis, as well.

Concerning anticipation, in line with the canonical view (compare Section 4.1.1) we expected the gesture to occur well before the initial linguistic token of the utterance. This means, that the start time of the preparation phase would come before the start time of the first word. We will begin by taking a classifying perspective on our data. The following categories seemed to be useful in approaching the phenomenon of gesture anticipation:

quite early: The gesture begins at least one second sooner than the first linguistic segment of the utterance.

early ("canonical"): The gesture starts less than one second earlier, but still before the first linguistic segment of the utterance. This is intended to capture McNeill's claim that a preparation usually *anticipates* the linguistic material *by a brief interval*.

late: The gesture begins at the same time or up to one second later than the first linguistic segment of the utterance.

quite late: The gesture starts more than one second later than the first linguistic segment of the utterance.

The results of employing this categorization system are displayed in Table 4.2. Given the original formulation of gesture anticipation in section 4.1.1, one seems to be free to conceive of it as a breakable constraint, as a sort of default which is right for the majority of cases. Our results are compatible with an interpretation like this.

occurrence of gesture	no. of tokens
quite early	1
early ("canonical")	16
late	4
quite late	4

Table 4.2: Results for anticipation in pre-test "three"

[6] (Kita et al., 1998) approaches the topic of movement phase classification solely on grounds of kinematic properties. This approach is attractive with respect to automatic recognition and delivers good inter-coder reliability ratings. See also Dell (1977), who first classified motion in terms of effort.

However, when we turn to a quantitative analysis, things begin to look more ambiguous. Table 4.3 below displays results for several measures of central tendency and dispersion, which have been calculated with respect to the same data.[7] The results have to be read as follows: among measures of central tendency, a negative sign means negative precedence for the gesture, i.e., the preparation phase starts after the onset of the first linguistic token of the multi-modal utterance. Therefore, the calculated arithmetic mean of -0.1712 comes as quite a surprise against the background of the occurrence class frequencies laid out in Table 4.2, given the predictions of the canonical view.

It is, however, less surprising once one takes into account the enormous temporal range of the data, with values spanning from -4.28 to 1.04 (seconds of gesture precedence)—corresponding to an overall range of 5.32. With a median of 0.12, modes of 0.08 and 0.36, and an interquartile range of "only" 0.60 (q_1 = -0.24, q_2 = 0.12, and q_3 = 0.36), it is clear that it must be mostly owing to negative runaway values (compiled under the heading of "quite late" performances in Table 4.2) that there is a standard deviation of 1.1058.

The latter is a comparatively large value, given the respective measures of central tendency. We will return to our findings pertaining to anticipation in perspective at the end of this section.

measure of central tendency	result	measure of dispersion	result
arithmetic mean	-0.1712	standard deviation	1.1058
median	0.12	interquartile range	0.60
mode(s)	0.08; 0.36	range	5.32

Table 4.3: Statistical results for anticipation in pre-test "three"

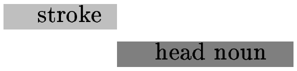
Turning to rules of *synchrony*, we will mainly be concerned with semantic synchrony here. Firstly, it has to be noted that McNeill's way of formulating this point (compare the quotes in Section 4.1.1) is problematic. This is because he uses the rather vague term "idea unit". McNeill's pertinent explications are not of much help when it comes to operationalizing his claims in a precise way. Our linguistic data consist of more or less complex nominal phrases that do get used in context to realize indirect speech acts of a directive illocutionary point (Searle and Vanderveken, 1985) on the level of pragmatics. Taking pragmatic synchrony for granted, it is still not clear which meaning is contributed by the deictic gesture and which linguistic package does thereby get complemented—in a more narrow sense of the word. We shall therefore attempt to cash out different interpretations of "idea unit". We will try a rather strict approach first, since, if successful, this would strenghten a more informative and therefore more interesting version of the rule of semantic synchrony. Therefore, we elaborate in the following on the hypothesis that only part of the nominal phrase is being meaning-complemented and synchronized with.

From the perspective of our underlying linguistic formalism (Sag and Wasow, 1999), signs are understood as multi-dimensional entities, i.e., to represent information on the phonological, syntactic and semantic level simultaneously. Nominal phrases, accordingly, are traditionally analyzed as being headed by their noun constituents. This means that in traditional terms the noun is conceived of as the most important daughter constituent ("head daugh-

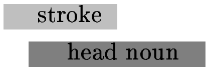
[7] Remember that the range of available data is too small for analytical statistics. We employ descriptive statistics here, because this move will be helpful in discovering hidden complexities.

ter") providing most of the information—or rather, the most important information—that gets projected to the representational level of the phrase.[8] Being arguably the most important element of the phrase, the noun looks like a good candidate for the linguistic part of the "idea unit". With the original formulation of semantic synchrony in the back of our heads, we will analyze our data with respect to the following categories:

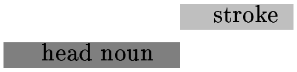
before head noun: the stroke ends before the start of the head noun[9].



on head noun: the stroke shares some arbitrarily long stretch of time with the head noun.



after head noun: the stroke starts after the end of the head noun[9].



Since we are looking at relationships between time intervals here, more formally minded definitions might make use of Allen's interval-based time logic (Allen, 1984). However, we will not do so, since the given definitions do suffice with respect to our rather modest aims here.[10] Results of employing the above categorization scheme are displayed in Table 4.4.[11]

Occurrence of gesture stroke	No. of Tokens
before head noun	5
on head noun ("canonical"?)	10
after head noun	6
not categorizable	4

Table 4.4: Synchrony in pre-test "number three"

At the very least, these results show that semantic synchrony, under the interpretation of "idea unit" applied here, cannot hold with respect to the underlying data. Although many gestural strokes can be seen to be homing in "on the head noun", more than half of the occurrences are observed strictly before or strictly after it. Generally speaking, gestural strokes could be observed to occur at strikingly different start positions with respect to their affiliated linguistic material, e.g. on preceding adjective modifiers or on post-poned modifiers as prepositional phrases or relative clauses. Furthermore, strokes were observed as having

[8] Hence the name: noun phrase. Note, that there are indeed other options available to the linguist, e.g., presuming that the determiner functions as head and therefore switch to speaking about determiner phrases. This objection, however, seems to have no force against the direction of our argument to come.

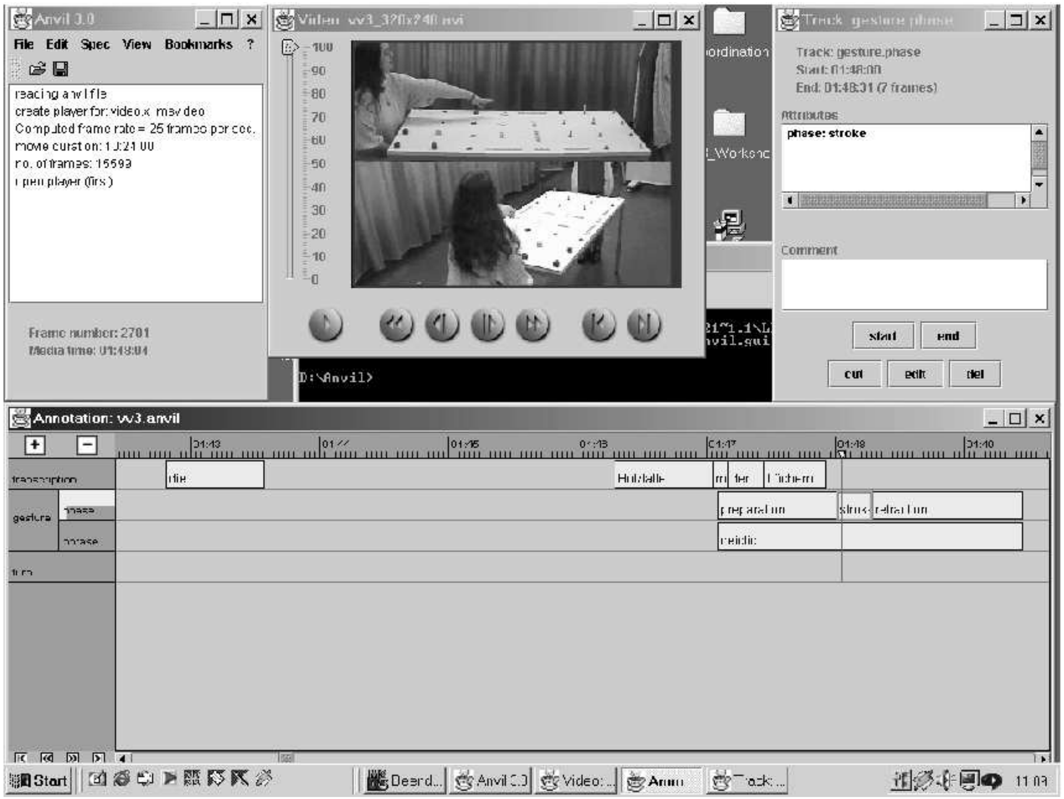
[9] There need not be an immediate succession of stroke and corresponding linguistic construction. A "slight offset" (to be judged case-wise on pragmatic grounds) is admissible.

[10] What is more, the definitions given above can be translated more or less directly into the algorithm format exploited in our XSLT scripts for means of automated analysis/classification.

[11] Cases compiled under "not categorizable" subsume those, where there is no head noun to be found on the linguistic surface, possibly in virtue of being an elliptical utterance or due to a performance error.

considerably different duration in time: with an arithmetic mean of 0.6176, respective values range from 0.08 to 1.60 seconds.

In the face of these negative results, the strict interpretation of "idea unit" has to be given up in order to gain a better fit with the data. For example, one might think of projections of type $\overline{N}$—i.e. configurations of head (common) nouns and their respective modifiers/complements, not encompassing the determiner—to provide the right level of abstraction for the linguistic part of the "idea unit" package. One might finally fall back on the whole noun phrase, which, however, provides the least interesting interpretation of "idea unit". We will not do so here. Instead, we shall exemplify a class of phenomena, whose observation made us think of whether there is a point in pursuing any of the afore mentioned strategies (and, correspondingly, in a strict rule of semantic synchrony).



```
reading anvil file
create player for: video.x msvideo
Computed frame rate = 25 frames per sec.
movie duration: 10:24.00
no. of frames: 15593
open player (first)
```

Frame number: 2701
Media time: 01:48:04

```
D:\Anvil>
```

Track: gesture.phase
Start: 01:48:00
End: 01:48:31 (7 frames)

Attributes

phase: stroke

Comment

Figure 4.3: Deictic stroke in post position

The annotation screenshot in Figure 4.3 depicts a stroke occurrence, where the stroke doesn not only follow the head noun (compare the criterion for category "after head noun" above), but all of the linguistic entities of the utterance. We will subsume such cases under the heading of *post-position strokes*. Obviously, they are incompatible with any interpretation of semantic synchrony. We readily admit that there is only one clear-cut case of a post-position stroke in our data basis. Nevertheless, as can be seen below in Section 4.1.2, there is a reasonably large set of cases, where the gestural stroke starts well after the start of the last linguistic token of the utterance.[12] Some of these tokens come *very* close to the strict criteria

[12] The entries under category "final" in Table 4.5 below are points in case.

for the sort of behavior depicted in Figure 4.3, so it seems that we are hunting for something real here.

As has been hinted at above, we will not have much to say concerning phonological synchrony, simply because we have not conducted a detailed analysis yet. Nevertheless, the annotated data available suffice to show that there are utterance tokens that obviously contradict the rule of phonological synchrony as formulated by McNeill. The example of a *post position stroke* depicted above is a clear point in case. However, similiar results can reasonably be expected to hold among those cases mentioned above that fall short ahead of the criteria for "post position strokes".

To summarize and elaborate on our findings with respect to the canonical view: first, our results can be seen to support a weak version of anticipation, which seems to be broadly in line with McNeill's perspective. Second, there are severe problems with gestural synchrony: under a narrow interpretation of "idea unit", semantic synchrony does not hold. Furthermore, part of the data seem to contradict all possible interpretations of "idea unit". However, a weaker version of synchrony—compare the default-like status of anticipation presumed here—could find our support.

Anyway, there seems to be sort of an inner incoherence in McNeill's way of stating the rules of anticipation and synchrony: if anticipation is formulated as a constraint that may be overridden from time to time, and, furthermore, anticipation can be seen to function as a prerequisite for synchrony (at least with respect to the majority of cases) then it seems to be at least unlikely that synchrony should hold for all cases.

This notwithstanding, more research will be necessary to show whether our suspicions can be corroborated. The jury is still out on the question whether the canonical view can be extended to task-oriented settings in general.

Note, that there is some debate in the pertinent literature[13] (e.g. (Nobe, 1996), (Levelt et al., 1985) and (de Ruiter, 1998)) with respect to (phonological) synchrony: the majority of the respective findings seem to strenghten a rule comparable to the one given by McNeill. However, the degree of synchrony remains a matter of debate. Our findings should be considered within the framework of this perspective.

Preliminaries to a Linguistic Account: The Problem of Temporal Concurrency

As has been noted above, the main theoretical goal of our project is the development of a representational account of multi-modal directives comprising speech and pointing, as can be observed in task-oriented dialog. Since we are claiming that deictic gestures have a status on a par with linguistic signs, one of the first questions to be answered can be put as follows: "How is linguistic and gestural information to be represented in one format?" With our theorizing being rooted in the methodological framework of linguistics, we have to provide some sort of surface symbol string as input for our apparatus to operate upon. We will construe it as corresponding to an abstract communication channel, one that is neutral with respect to the actual modalities exploited. Occurrences of deictic gesture strokes will be indicated by using the symbol "↘", see (Rieser, 2001). Taking these abstractions for granted, one has to think about how to realize the mapping from concurrently time-smeared entities onto a linear

[13] We would like to thank Timo Sowa, University of Bielefeld, for help and information regarding the literature.

succession of symbols[14].

Since we are dealing with concurrent time intervals, one might try to state abstract mapping rules, whereby the relationships among the relevant entities involved should serve as triggering conditions for the application of respective rules. The language of Allen's interval-based time logic (Allen, 1984) could be used in order to state the corresponding conditions in a sufficiently precise way. However, such an approach cannot be realized straightforwardly, since entities on the gestural channel may be combined with an arbitrary number of entities on the linguistic channel, and thereby the number of possible triggering conditions for mapping rules grows infinitely large. Even if one would find a working finite sub-set of those rules, it would be far from obvious which principles should guide the mapping process, since there would be an enormous wealth of structural configurations to be accounted for.

As a workable approach to bring about the mapping in a systematic way, we make use of *anchors* as "representatives" for respective time intervals. For every event on the gestural and linguistic channel, we systematically provide a single time-stamp. Given such anchors, mapping is defined to be isomorphic with respect to the ranking of anchors on the time-scale. Now, there are at least three systematic solutions to the problem of deciding on representative anchors: we can go for the start time, the end time, or we can compare with respect to the arithmetic mean of both. Note that slightly different mappings will result, depending on the implemented strategy. Here, we will exemplify results according to the first strategy:

position of "↘" in surface string	No. of tokens
first	2
final	9
in between	14

Table 4.5: Position of ↘ in linear input string representation

With respect to the terminology introduced above, "first" means that the anchor of a respective gestural stroke is ranked in front of the anchor of the first linguistic element of the utterance. At the other extreme, "final" means that the anchor of the gestural stroke is ranked behind the anchor of the last linguistic entity. Simple examples from our data are given below: (4.1) exemplifies "final", (4.2) and (4.3) are examples for "in between", and (4.4) belongs to the category "first".

(4.1) die grüne Schraube ↘
the green bolt ↘

(4.2) die rote ↘ Schraube
the red ↘ bolt

(4.3) die ↘ gelbe Schraube
the ↘ yellow bolt

[14] This is a non-trivial problem and one of a sort that linguists rarely, if at all think about. From the perspective of generative grammar, time figures as an abstract factor that does not play a prominent role in respective analyses. At best, on a gross scale, temporal succession can be seen to be encoded implicitly in form of linear succession of symbols on the level of the surface string.

(4.4) ↘ die Leiste
↘ the bar

An important aspect of our theoretical work, which is rooted in a constraint-based lexicalist species of generative grammar (Sag and Wasow, 1999) consists in the specification of grammars for this abstract representation format and has to provide syntactic and semantic analyses as appropriate with respect to the empirical basis (Kühnlein et al., 2003).

4.2 Spatial Phenomena

Figure 2.2 shows the distribution of the objects we presented to our subjects (Ss) for one of the possible clusterings. The Ss were located at the table, as shown in Figure 4.4 for convenience.

While the instructor sat on a chair close to the table, the constructor had to be able to move around in order to be able to touch and lift the objects which were indicated by the instructor. This led to the following situation: some of the objects (the ones close to the left edge of the table in Figure 4.4) were within reach of the instructor's arms. The remainder was not within reach of the instructor, though clearly for the constructor.

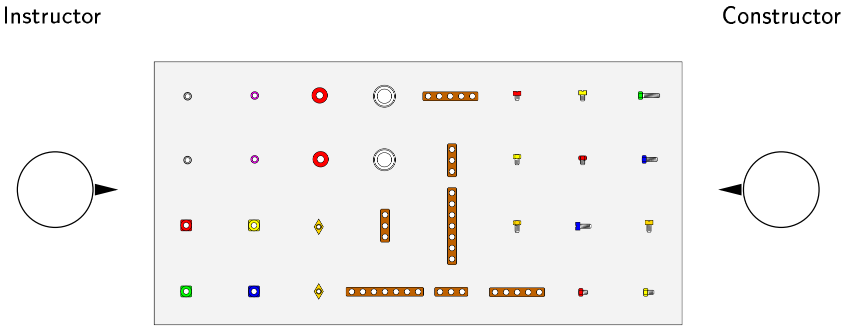


Figure 4.4: In the studies, instructor and constructor were located opposite each other, both facing the table with the clustered objects as pictured in Figure 2.2

4.2.1 Proximal Objects

Interestingly, the instructor regularly refused to point towards objects within reach from her/his location. Instructors preferred rather to simply gaze at the objects while uttering the instruction, and one of them even lifted an object herself. This suggests that people tend not to use pointing gestures when they can resolve the reference of their utterances by easier means, e.g., gaze. When subjects pointed at objects in their immediate vicinity, they often did so with an expression of amusement on their faces, or giggling. This might indicate that the Ss thought of such an action as inappropriate under such circumstances. The verbal utterances when reference to near objects was made tended to be short and simple, e.g., "That".

4.2.2 Distal Objects

On the assumption that pointing by finger is a somehow fuzzy and not sharply determined physical action, we had predicted that there is a limit to the "resolution capacity" of the pointing finger. This prediction is supported by the fact that, as we observed, Ss did not rely on the use of pointing to objects close to the opposite side of the table[15]. This is supported by the fact that they tended to use complicated descriptions, with a strong increase in length as compared to the utterances used in the case previously mentioned, in just half of the cases accompanied by pointing gestures. A typical utterance for this variety is

(1) Take the yellow bolt at the corner.

Some of the utterances were in themselves already sufficient to identify the respective objects laid out on the table, as in the case given: there was a unique yellow bolt at a corner of the table. It seemed that here gestures were only used as an auxiliary device to indicate the direction in which constructors had to search for the objects described.

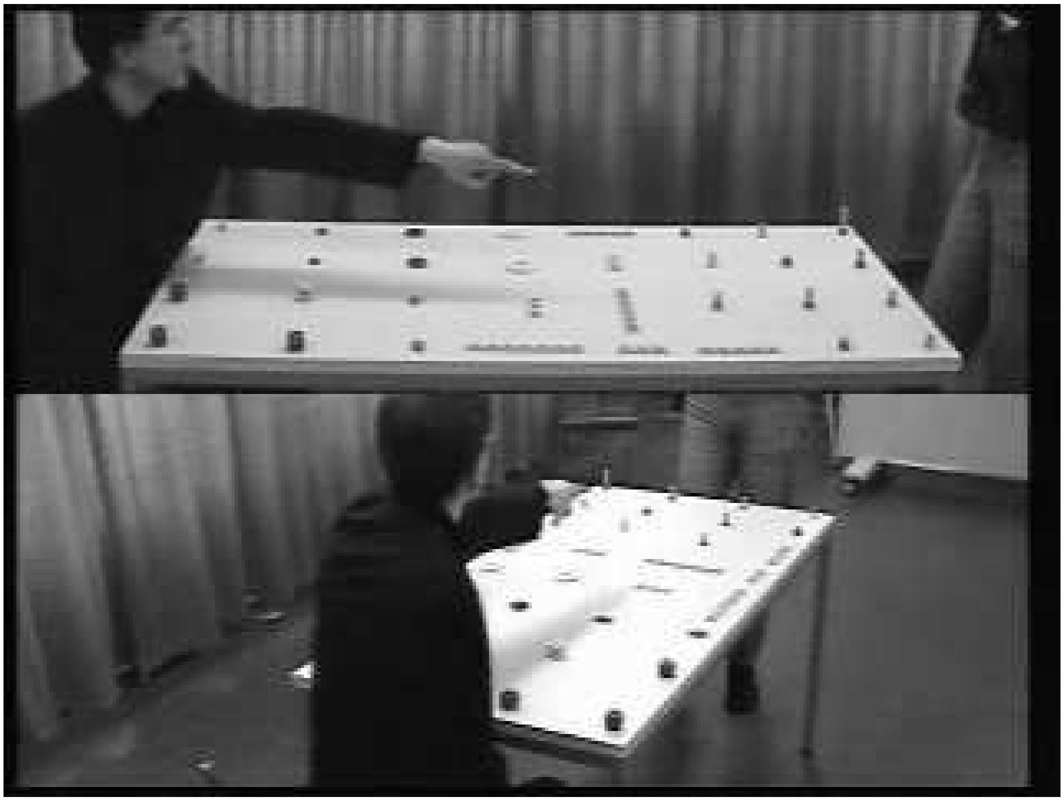


Figure 4.5: This is a case where pointing is used *in addition to* a comparatively long description ("Die rote Schraube"/"The red bolt").

Figure 4.5 shows an instance where the instructor indeed uttered a comparatively long description of the intended object, accompanied by a pointing gesture. These cases, however,

[15] Piwek and Beun (2001) observe changing use in proximate/distal pronouns depending on the spatial conditions. They also report on changing behavior concerning the use of post-modifiers. Their findings will be tested for in our videotapes and the results reported in some following report.

made up only half of our data concerning distant pointing.

This conjecture concerning distant objects can be sustained by the following observation: there were cases in which instructors pointed at distant objects, using only short descriptions like “That bolt”. In these cases constructors had greater difficulties to identify the correct objects. So it seems that the gesture alone did not suffice for the purpose at hand. On the other hand, there were cases where instructors did not use a pointing gesture at all when they referred to distant objects verbally. These cases did not show any difference in terms of success to the cases where the utterances were accompanied by gestures.

4.2.3 Mid-ranged Objects

Most occurrences of pointing gestures involving fingers were observed to be directed towards the mid area of the table. The verbal utterances that accompanied them were of medium length, and did often not allow identifing the objects if they were not accompanied by gestures. Gaze alone obviously did not suffice to identify the objects in that area, either. So it seems that pointing gestures serve primarily as a means to identify objects that are in a certain distance from the person pointing, in our case the instructor. Figure 4.6 gives an impression of the structure that is thus imposed on the domain that is constituted by the objects on the table.

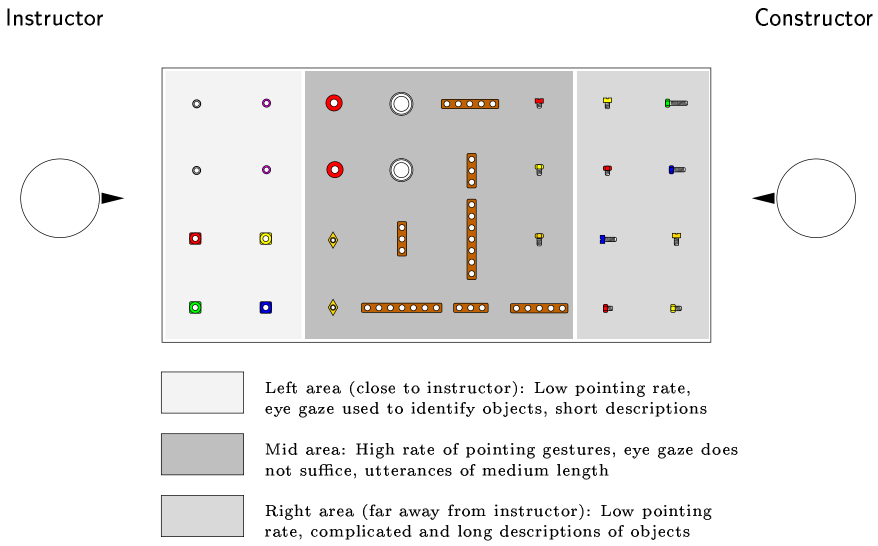

Instructor
Constructor
Left area (close to instructor): Low pointing rate, eye gaze used to identify objects, short descriptions
Mid area: High rate of pointing gestures, eye gaze does not suffice, utterances of medium length
Right area (far away from instructor): Low pointing rate, complicated and long descriptions of objects

Figure 4.6: The structure imposed on the domain, according to the occurrence and success of pointing gestures.

Based on these observations, the hypothesis can be formulated that pointing gestures can be used to identify objects only in case certain physical conditions obtain among the objects which serve as alternatives for reference. One of the conditions that has to be met, more specifically, is that a certain subjective density may not be exceeded. What is meant by subjective density is the density as perceived by the instructor. Note that in our setting the real density was the same everywhere on the table. We had 32 objects laid out on an area of 70×140 cm^2, corresponding to $\approx .003$ objects/cm^2 throughout the total area. Things look different, however, from the observer’s point of view. We refer to Figure 4.7 to point out how the scene is perceived from one of the subjects’ perspective. Figure 4.7 is, of course, an enriched version of Figure 2.1, page 5. What we added are some lines to bring out the following points:

4.2.4 Subjective Density

The objects closest to the instructor appear to be arranged with less density than those farer away from him subjectively. It is also obvious that the instructor can easily interfere with the objects in her/his vicinity, but that this is limited: the pointing hand will not reach further into the table than about 60 cm as long as the instructor is in an upright position. Indeed, there are a few examples where the instructor not only does stretch out the arm very far, but bends over the table to come closer to objects pointed at.

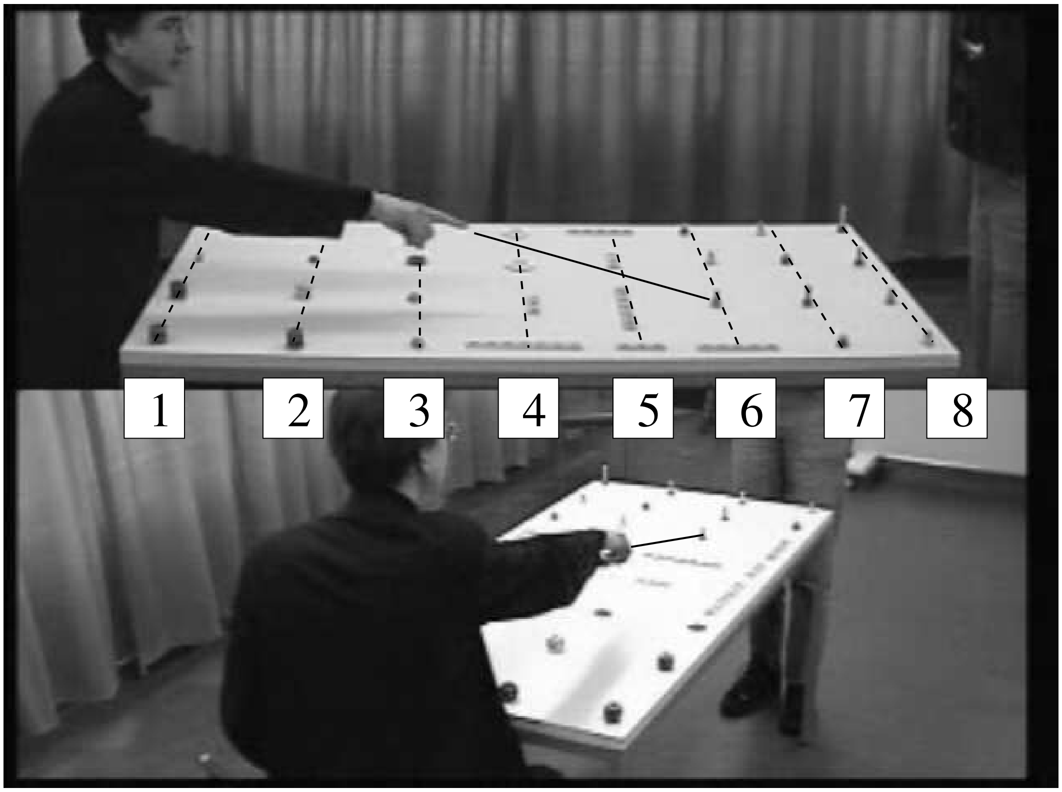


Figure 4.7: A sample still from pre-test 5.

The subjective density can be calculated in the following way: Let h be the height of the knuckle of index finger over the table; let d be the horizontal distance of the index fingers knuckle to the object pointed at. Then the ratio d/h is the tan of the angle α that is included between the index finger and the vertical line through the knuckle. This means that the finger can point with, say, 45° when h=10 cm and d=10 cm, but also with multiple other heights and distances. This is due to the fact that in all cases in which the ratio $d/h = 1$, the arctan (the inverse of the tan) is 45°. The following holds:

$$\begin{aligned} \tfrac{d}{h} &= \tan\phi \\ \arctan\tfrac{d}{h} &= \phi \end{aligned} \tag{4.5}$$

d=20 cm and h=10 cm accordingly gives an angle of 63.4°, measured against the vertical line through the knuckle. Notice that tan, of course, is not a linear function: Double distance and equal height does not mean that the angle is doubled etc. Rather, the limit function $\lim_{d \to \infty} \arctan d = 90^{\circ}$, which corresponds to a right angle against the vertical during pointing to an object at an arbitrarily large distance. As a corollary, this also implies that the angles between lines pointing to linearly increasing distant objects differ by ever smaller amounts.

The characteristics of arctan are responsible for the different impact attributable to distances. E.g., pointing to an object that is 50 cm away (with h=10 cm) takes place at an angle of 78.7°, while pointing to an object 10 cm further away corresponds to an angle of 80.5°. Compare this difference of 1.8° with the difference calculated above for 10 cm and 20 cm, which is 18.4°.[16] This indeed is the interesting interaction when it comes to calculating the subjective density.

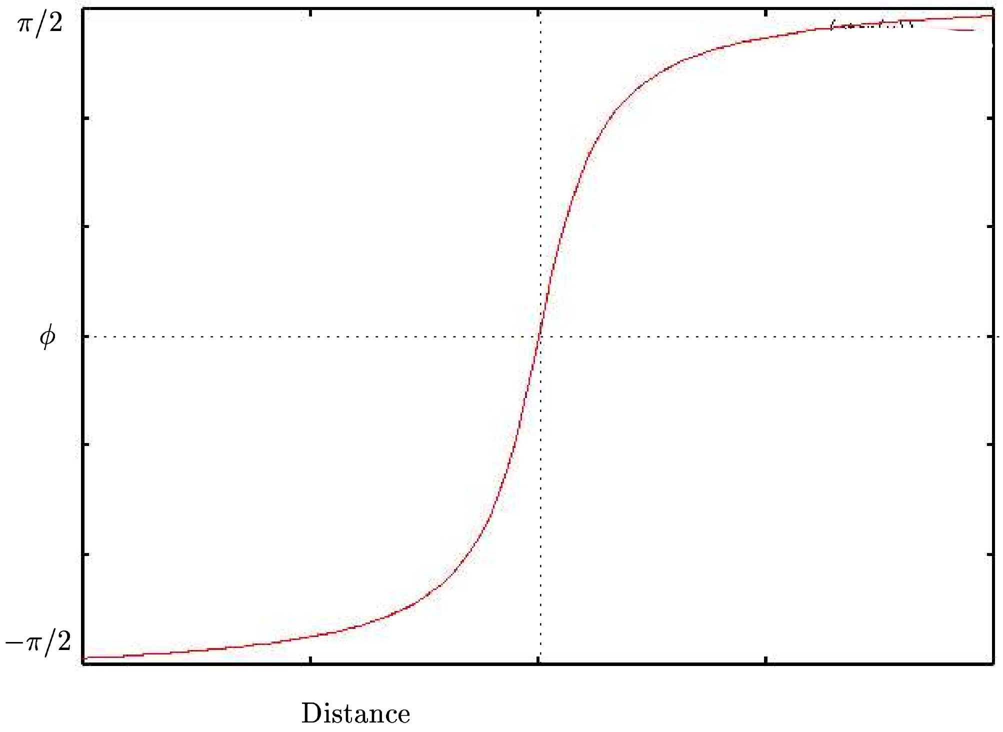


Figure 4.8: The characteristics of *arctan*: linear increase in distance and constant height lead to ever smaller differences in angle

The setting we presented to our subjects consisted of objects that were distributed regularly over a table, the mean distances were ≈17 cm. If the knuckle of the index finger reaches approximately 40 cm into the area of the table in the average, and its height is 20 cm over the table average, this means that the objects at the far end (row 8 in Figure 4.7) are at a

[16] Another case of the impact of the behavior of arctan comes out when looking at the variation of the h-value: when d is kept constant, larger h means a smaller angle. Again, the dependence is non-linear. Pointing to an object that is 10 cm away from a height of 20 cm results in an angle of merely 26.6°. d=30 cm corresponds to an angle of 18.4°, etc. The limit of the function $\lim_{h \to \infty} \arctan \frac{1}{h} = 0^{\circ}$, which corresponds to a zero angle against the vertical during pointing to an object at an arbitrarily large height.

distance of 90 cm. (The position of the objects is 10 cm from the edge of the table.) This corresponds to an angle of 77.5°. The next row (7) of objects is 17 cm closer to the instructor, which means that the angle is 72.3°. The fact that subjects chose to use longer descriptions for the identification of these objects suggests that an angular difference of 5° is too small to resolve the intended object. Yet, the next row (6) is usually described in a complicated way. It is separated by an angle of 5.8° from the even closer row. So, obviously 6° is too little a difference to be reliable for identification. The next row (5) of objects is only 39 cm away from the knuckle of the finger. The corresponding angle, 55.4°, is roughly 8° smaller than the one that marked the previous difference. And usually this seems to be regarded as a sufficient difference, as pointing takes place here with shorter descriptions. It seems that the resolution of the pointing finger lies somewhere between 6° and 12°[17].

The fact that the hand reaches approximately 40 cm into the area of the table explains only half way why the objects in the first two rows are not as frequently pointed at as the objects in the next four rows. Consider, however, that the instructor's *eyes* are further away from the table than his hand. If the same relations held for "pointing" with the eyes as for pointing with the fingers, it could be explained why within this distance referring can be successful without the use of the index fingers[18].

Suppose the instructor's eyes are approximately 40 cm above the table and in line with the tables edge. Then an object that is at a distance of 10 cm from the edge is looked at with an angle of 14° as measured against the vertical line. The next object, being 27 cm away, is then 20° away from the first row, clearly within the expected range. The third row is only 13.7° further away from that row, coming close to the distance being problematic.

It may now be clear, why it did not matter much that the subjects moved the objects a little during the studies, as mentioned in subsection 2.2: the differences that occurred in the course of this were too small to be relevant as long as we observe angles on this magnitude.

These observations fit well with findings that are summarized in (Oviatt, 1999, 78). Oviatt reports that multimodal utterances are shorter and syntactically simpler than unimodal (spoken) utterances. Assume that utterances lose their multimodal character when pointing can not be exploited owing to high subjective density. It is consistent then with Oviatt's claims that the linguistic material under this condition becomes much more complicated: Oviatt (1999), *loc. cit.*, states that

> [w]hen free to interact multimodally, users selectively eliminate many linguistic complexities. As illustrated here, they prefer not to speak error-prone spatial location descriptions [...] if a more compact and accurate alternative is available, such as pen input.

It is not quite clear what Oviatt has in mind with the expression "compact and accurate". Probably she means "compact" in the sense of "less time consuming", and "accurate" in the sense of "uniquely identifying".

[17] There is an interesting connection to a dissertation that is currently written at the Univ. Birmingham by Kirsty Crombie-Smith. She observes that deaf can resolve movements of fingers as little as 3°. This is, of course, half the size of the lower bound of the interval we are assuming. An object lying at the center of the area circumscribed by the 6°-angle would be positioned at the edge of it after a movement of 3°. The findings can, however, be compared only with difficulty, as the conditions under which they were obtained are too different. Crombie-Smith reported her findings at the Gesture Conference at Genova in 2003.

[18] This might account for the findings in (Steininger et al., 2001), where Ss are reported not to have used *interactional* (= pointing, iconic etc.) gestures at all in an Wizard-of-Oz scenario. The screen that was used as a display in their setting was always *within* reach of the agent.

Note that Oviatt does not mention a connection with spatial conditions like the subjective density we found.

4.2.5 The Conceptualization of the Pointing Space

To sum up, it seems to be the case that the space in which the interaction between instructor, constructor and the objects takes place is conceptually structured by the participants. Given the experimental setup, it is evident from the subjects' behavior that pointing in the close vicinity is commonly regarded as superfluous and a kind of strange thing to do. Subjects rely on the possibility to identify objects within a certain distance from instructor by ways of eye-gaze. Outside that range, gesture has a varying degree of reliability for object identification. This suggests that a valid semantic model for deictic gestures has to reflect the structure of the agents' environments during their interaction. In the present case at least, not only have the perspectives of the agents to be captured, but also the distribution of the objects in the common physical space.

Chapter 5

Concluding Remarks & Pointers to Future and Ongoing Work

We reported on empirical work that was done with respect to the way pointing gestures are used in our empirical setting. We found that the temporal relations exhibited by pointing gestures in comparison to speech are more complex than the literature reporting the relations suggests. This will be reflected upon in a follow-up report that is concerned with theoretical issues, especially the linguistic modelling of speech/gesture units.

Further, we brought up the topic of spatial conditions that are relevant for pointing and that are rarely studied. Our first observation is that subjects refuse to use their fingers to point in their vicinity, preferring to point by gaze. There will be a follow-up study that deals with the relevance of gaze direction. A second observation is that pointing to distal objects seems to work only if either it is accompanied by a complex spoken utterance or there is only one object within an angle of 6^o–12^o.

The results suggest that true computational models for pointing gestures have to respect two dimensions of expression: the first is, of course, the temporal occurrence of the stroke of the gesture. True models have to take into account that there is much more flexibility in the temporal relations than the canonical view would lead to expect. The second dimension concerns the spatial relations that hold during the pointing gesture. Both dimensions have already been included in the implementation of an embodied conversational agent called MAX, cf. (Kopp and Wachsmuth, 1999). MAX can now be parameterized pertaining temporal stroke precedence and along the dimension of the pointing cone, which reflects the effects of subjective density. The standard setting for the pointing cone is 8^o. Recent work is addressed in (Kranstedt et al., to appear).

We settled on an annotation format that is compatible with the one used by the computer simulation part of our project in order to control MAX. It was possible to automatically translate annotations of videos into MURML[1]. This enables the use of MAX for experimental purposes, where subjects conduct construction dialogues with it. We can use MAX to replicate actual human-human studies with an artificial instructor whose pointing behaviour can easily be modified by re-setting parameters. We can thus conduct empirical studies in a way that were not possible before. Besides, the execution of the MURML code is an immediate opportunity to verify the annotations.

The annotation of subjective density should become a standard for corpora that are in-

[1] A language realized in XML and based on HamNoSys and McNeill's framework used to control MAX.

tended for research into pointing gestures. We separately fed the data into the scene description for Max, but it is clear that this is only a preliminary solution. An annotation standard will accordingly be proposed at some Message Understanding Conference (MUC) or similar conference in the future. Two solutions are possible in principle: (i) Description of the geometrical dimensions of the (virtual) setting in a separate part of the dialog annotation plus description of the finger positions at each part of the corpus. (ii) Description of the finger position at each part of the corpus plus description of the position of objects within the pointing cone. For our purpose, (i) will be the better solution, as Max needs a Vrml description of the geometrical data anyway. All aspects concerning annotation (including the spatial and temporal dimensions) will be treated in greater detail in a follow-up to this report.

A side effect of the empirical studies conducted with Max will be that we can fine-tune the conversational agent by evaluating the behavioural data we obtain against those collected in human-human studies.

Acknowledgements

The work leading to this paper was done within the framework of the SFB 360. We want to thank our colleagues in the SFB, especially Andy Lücking, Manja Nimke, and Hannes Rieser, who helped conducting the studies and with whom we could discuss all the aspects of the paper. Stefan Kopp, Alfred Kranstedt, Timo Sowa and Ipke Wachsmuth helped improving the paper by providing us with all the technical details about Max and Murml. Thanks go also to Claudia Sassen, Univ. Dortmund, for English corrections and help with all questions of style. Without her aid many of the formulations in this text would have been completely cryptic.

Bibliography

Allen, J. F. (1984). Towards a General Theory of Action and Time. *Artificial Intelligence*, 23:123–54.

Atkinson, J. M. and Heritage, J., editors (1984). *Structures of Social Action—Studies in Conversation Analysis*. Cambridge UP.

Bernsen, N. O., Dybkjær, L., and Kolodnytsky, M. (2002). The NITE Workbench—a Tool for Annotation of Natural Interactivity and Multimodal Data. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*.

Bernsen, N. O. and Stock, O., editors (2001). *International Workshop on Information Presentation and Natural Multimodal Dialogue*—IPNMD 2001. ITC-irst.

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., and Liberman, M. (2000). ATLAS: A flexible and extensible architecture for linguistic annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pages 1699–706.

Bird, S. and Liberman, M. (1999). Annotation Graphs as a Framework for Multidimensional Linguistic Data Analysis. In *Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop*. LDC.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1, 2):23–60.

Carletta, J., Isard, A., Isard, S., Kowtko, J., and Doherty-Sneddon, G. (1996). HCRC dialogue structure coding manual. Technical Report TR-82, HCRC. http://www.hcrc.ed.ac.uk/publications/tr-82.ps.gz.

Carletta, J., McKelvie, D., Isard, A., Mengel, A., Klein, M., and Møller, M. B. (2002). A generic approach to software support for linguistic annotation using XML. In Sampson, G. and McCarthy, D., editors, *Readings in Corpus Linguistics*. Continuum.

de Ruiter, J.-P. d. (1998). *Gesture and Speech Production*. PhD thesis, Katholieke Universiteit Nijmegen. MPI Series in Psycholinguistics.

Dell, C. (1977). *A Primer for Movement Description using Effort-Shape and Supplementary Concepts*. Dance Notation Bureau Press, revised edition.

Dybkjær, L. (2000). MATE Final Report. Technical Report MATE Deliverable D6.2, MATE.

Dybkjær, L. and Bernsen, N. O. (2002). Natural interactivity resources—data, annotation schemes and tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*.

Dybkjær, L., Bernsen, N. O., Dybkjær, H., McKelvie, D., and Mengel, A. (1998). The MATE Markup Framework. Technical Report MATE Deliverable D1.2, MATE.

Hanfmann, E. and Vakar, G., editors (1962). *Vygotsky, Lev Semenovich: Thought and Language*. MIT Press. Transl. by the editors.

Harold, E. R. and Means, S. (2001). *XML in a Nutshell*. O'Reilly.

Kendon, A. (1972). Some Relationships between Body Motion and Speech: An Analysis of an Example. in: (Siegman and Pope, 1972).

Kendon, A. (1980). Gesticulation and Speech: Two Aspects of the Process of Utterance. in: (Key, 1980).

Key, M. R., editor (1980). *The Relationship of Verbal and Nonverbal Communication*, volume 25 of *Contributions to the Sociology of Language*. Mouton.

Kita, S. (1990). The temporal relationship between gesture and speech: A study of Japanese-English bilinguals. Master's thesis, Department of Psychology, University of Chicago.

Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. *Linguistics*, 35:379–415.

Kita, S., v. Gijn, I., and vd. Hulst, H. (2000). Gesture Encoding. Technical report, MPI Nijmegen.

Kita, S., van Gijn, I., and van der Hulst, H. (1998). Movement Phases in Sings and Co-speech Gestures, and Their Transcription by Human Coders. in: (Wachsmuth and Fröhlich, 1998).

Kopp, S. and Wachsmuth, I. (1999). Natural timing in coverbal gesture of an articulated figure. Working notes at the workshop 'Communicative Agents' at "Autonomous Agents 1999", Seattle.

Kranstedt, A., Kopp, S., and Wachsmuth, I. (2002a). MURML: A Multimodal Utterance Representation Language for Conversational Agents. Technical Report 05/2002, SFB 360. also as (Kranstedt et al., 2002b).

Kranstedt, A., Kopp, S., and Wachsmuth, I. (2002b). MURML: A Multimodal Utterance Representation Language for Conversational Agents. In *Proceedings of the Workshop Embodied Conversational Agents - let's specify and evaluate them! held at First Int. Joint Conference on Autonomous Agents & Multi-Agent Systems*.

Kranstedt, A., Kühnlein, P., and Wachsmuth, I. (to appear). Deixis in Multimodal Human-Computer Interaction. In *Proceedings of the 5th Gesture Workshop*, LNAI, Genova. Infomus, Springer.

Kühnlein, P., Nimke, M., and Stegmann, J. (2003). Towards an HPSG-based Formalism for the Integration of Speech and Co-verbal Pointing. Online-proceedings of the Gesture Conference Austin (Texas). http://www.utexas.edu/coc/cms/International_House_of_Gestures/Conferences/Proceedings/Contents/List_of_Papers.html.

Levelt, W. J., Richardson, G., and La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, 24:133–64.

Lobin, H. (2000). *Informationsmodellierung in XML und SGML.* Springer.

Martell, C. (2002). FORM: An Extensible, Kinematically-based Gesture Annotation Scheme. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC).*

McKelvie, D., Isard, A., Mengel, A., Møller, M. B., Gross, M., and Klein, M. (2001). The MATE workbench—an annotation tool for XML coded speech corpora. *Speech Communication*, 33(1–2):97–112.

McNeill, D. (1992). *Hand and Mind—What Gestures Reveal about Thought.* Chicago UP.

McNeill, D., editor (2000). *Language and Gesture.* Cambridge UP.

Milde, J.-T. and Gut, U. (2001). The TASX-environment: An XML-based corpus database for time aligned language data. In *Proceedings of the IRCS Workshop on linguistic databases, Philadelphia.*

Mitkov, R. (2002). *Anaphora Resolution.* Longman.

Nobe, S. (1996). *Representational Gestures, Cognitive Rhythms, and Acoustic Aspects of Speech: A Network/Treshold Model of Gesture Production.* PhD thesis, University of Chicago, Department of Psychology.

Oviatt, S. L. (1999). Ten Myths of Multimodal Interaction. *Communications of the ACM*, 42(11):74–81.

Piwek, P. and Beun, R. J. (2001). Multimodal Referential Acts in a Dialogue Game: From Empirical Investigation to Algorithm. in: (Bernsen and Stock, 2001).

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., and Henning, J. (1989). *HamNoSys, Version 2.0—Hamburg Notation System for Sign Languages: An introductory guide*, volume 5 of *Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser.* Signum Seedorf.

Ray, E. T. (2001). *Learning XML–Guide to Creating Self-Describing Data.* O'Reilly.

Rieser, H. (2001). A Unified Account for Gesture Meaning and Expression Meaning in Simple Reference Games. Unpublished Manuscript.

Rimé, B. and Schiaratura, L. (1991). Speech and gestures. In Feldman, S. and Rimé, R., editors, *Fundamentals of nonverbal behavior*, pages 239–81. Cambridge UP.

Sag, I. and Wasow, T. (1999). *Syntactic Theory—A Formal Introduction.* CSLI.

Schegloff, E. A. (1984). On some gestures' relation to talk. in: (Atkinson and Heritage, 1984).

Searle, J. R. and Vanderveken, D. (1985). *Foundations of Illocutionary Logic.* Cambridge UP.

Siegman, A. W. and Pope, B., editors (1972). *Studies in Dyadic Communication.* Pergamon Press.

Sowa, T. and Wachsmuth, I. (2002). Interpretation of Shape-Related Iconic Gestures in Virtual Environments. In Wachsmuth, I. and Sowa, T., editors, *Gesture and Sign Language in Human-Computer Interaction*, number 2298 in Lecture Notes in Artificial Intelligence, pages 21–33. Springer.

Sperberg-McQueen, C. M. and Burnard, L., editors (1994). *Guidelines for Electronic Text Encoding and Interchange.* May. Text Encoding Initiative P3.

Steininger, S., Schiel, F., and Louka, K. (2001). Gestures During Overlapping Speech in Multimodal Human-Machine Dialogues. in: (Bernsen and Stock, 2001).

Thompson, H. and McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents. In *SGML Europe '97.* http://www.ltg.hcrc.ed.ac.uk/ dmck/sgml-europe-97.html.

Wachsmuth, I. and Fröhlich, M., editors (1998). *Gesture and Sign-Language in Human-Computer Interaction.* Springer Verlag. Proceedings of the International Gesture Workshop at Bielefeld, Germany, September 17–19, 1997.

Wittenburg, P., Levinson, S., Kita, S., and Brugman, H. (2002). Multimodal Annotations in Gesture and Sign Language Studies. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC).*