Assessing Reliability on Annotations (2): Statistical Results for the DEIKON Scheme

Andy Lücking Jens Stegmann
{andy.luecking,jens.stegmann}@uni-bielefeld.de

Collaborative Research Centre
SFB 360 "Situated Artificial Communicators"
B3 "DeiKon"

Abstract

This is the second part of a two-report mini-series focussing on issues in the evaluation of annotations. In this empirically-oriented report we lay out the documentation of the annotation scheme used in the DEIKON project, discuss the results obtained in a respective reliability study and conclude with some suggestions regarding forthcoming versions of the scheme. Relevant statistical background, theoretical considerations in reliability statistics and an evaluation of some pertaining approaches are given in the first, more theoretically-oriented report [Stegmann and Lücking, 2005]. The following points are dealt with in detail here: we describe the setting that was used to elicit the empirical data. The annotation scheme that is put to scrutiny is documented and exemplified. Aspects of our theoretical work in linguistics are mentioned *en passant*. Then we present, discuss, and interpret the actual results obtained for our scheme. We find a high degree of correlation on the exact placement of time-stretched entities (word and gesture phase boundaries), mildly good results pertaining to agreement concerning time-related categories that appeal to structural configurations (e. g. the position of a gesture with respect to the parts of accompanying speech), but rather weak agreement with respect to the determination of gesture function. Therefore, the results for time-based type-i data look more promising than those obtained for the more theoretically-framed type-ii categories. However, the type-i results must not be compared with the type-ii ones on superficial grounds, since the statistics are of a different kind (correlation *vs.* agreement, i. e. not chance-adjusted *vs.* chance-adjusted) and, hence, the results have to be interpreted in different terms, respectively. Finally, we discuss some issues in the future make-up of the annotation scheme with a focus on its dialogue parts. Our respective suggestions amount to a shift towards a more theory-oriented annotation.

Acknowledgements

We shall start this report by saying “thank you” to some very important persons. That is, we joyfully express our gratitude towards the following people:

Peter Kühnlein and Manja Nimke, who are former project members that have participated in the conduct of the original empirical studies, as well as in the discussions and decisions that helped to shape the very first version of the present annotation scheme.

Hannes Rieser, our project leader, who, besides having escorted earlier versions in the past, has pushed and carried the continuation and broadening of the annotation work with an emphasis on dialogue phenomena. Furthermore, he was our co-author on a workshop contribution that proved to be an early milestone on our way to this report (he humbly refused from being mentioned as an author of the present work). Last but not least, he was an invaluable aid with regards to corrections and proof-reading all parts of this document. This notwithstanding, of course, all the remaining errors and possible sources of misconceptions—it deems to us, there may be some—are truely and utterly ours.

We also wish to thank our DEIKON project partners in computer science/artificial intelligence, namely Ipke Wachsmuth, Stefan Kopp, Marc Latoschik, Timo Sowa, Alfred Kranstedt, and Thies Pfeiffer. Furthermore, we would like thank all members of the Collaborative Research Centre SFB 360 “Situierte Künstliche Kommunikatoren” at Bielefeld University, wherein our project is located and which has been a great place to work in.

Finally, we wish to express our deepest gratitude towards our families, and especially to Christiane and Karola. They patiently endured all inconveniences that arose during our work on this document. Without their support, this report could not have been completed.

Bielefeld, December 2005, Andy Lücking and Jens Stegmann.

[*Rosencrantz and Guildenstern pass their time by betting on the toss of a coin in the following manner. Guildenstern takes a coin out of his bag, spins it, lets it fall. Rosencrantz studies it, announces it as "heads" (as it happens) and puts it into his own bag. They have been doing this for some time and are witnesses of a highly improbably run of "heads" for ninety-two times in a row. Guildenstern, who is losing all the time, is well alive to the oddity of it. He is worried about the implications, not so much about the money he loses. We enter the dialogue midway (cf. the preface of the first report).*]

[. . .]

ROSENCRANTZ: Eighty-nine.

GUILDENSTERN: It must be indicative of something, besides the redistribution of wealth. List of possible explanations. One: I'm willing it. Inside where nothing shows, I am the essence of a man spinning double-headed coins, and betting against himself in private atonement for an unremembered past.

ROSENCRANTZ: Heads.

GUILDENSTERN: Two: time has stopped dead, and the single experience of one coin being spun once has been repeated ninety times. . . . On the whole, doubtful. Three: divine intervention, that is to say, a good turn from above concerning him, cf. children of Israel, or retribution from above concerning me, cf. Lot's wife. Four: a spectacular vindication of the principle that each individual coin spun individually is as likely to come down heads as tails and therefore should cause no surprise each individual time it does.

ROSENCRANTZ: I've never known anything like it!

GUILDENSTERN: Syllogism: one, probability is a factor which operates within natural forces. Two, probability is not operating as a factor. Three, we are now within un-, sub- or supernatural forces. Discuss. Not too heatedly.

ROSENCRANTZ: I'm sorry I—what's the matter with you?

GUILDENSTERN: The scientific approach to the examination of phenomena is a defence against the pure emotion of fear. Keep tight hold and continue while there's time. Now—counter to the previous syllogism: tricky one, follow me carefully, it may prove a comfort. If we postulate, and we just have, that within un-, sub- or supernatural forces *the probability is* that the law of probability will not operate as a factor, then we must accept that the probability of the *first* part will not operate as a factor, in which case the law of probability *will* operate as a factor within un-, sub- or supernatural forces. And since it obviously hasn't been doing so, we can take it that we are not held within un-, sub- or supernatural forces after all; in all probability, that is. Which is a great relief to me personally. . . . Which is all very well, except that—we have been spinning coins together since I don't know when, and in all that time (if it *is* all that time) I don't suppose either of us was more than a couple of gold pieces up or down. I hope that doesn't sound surprising because it's very unsurprisingness is something I am trying to keep hold of. The equanimity of your average tosser of coins depends upon a law, or rather a tendency, or let us say a probability, or at any rate a mathematically calculable chance, which ensures that he will not upset himself by loosing too much nor upset his opponent by winning too often. This made for a kind of harmony and a kind of confidence. It related the fortuitous and the ordained into a reassuring union which we recognized as nature. The sun came up about as often as it went down, in the long run, and a coin showed heads about as often as it showed tails. Then a messenger arrived. We had been sent for. Nothing else happened. Ninety-two coins spun consecutively have come down heads ninety-two consecutive times ... and for the last three minutes on the wind of a windless day I have heard the sound of drums and flute.

(from *Rosencrantz And Guildenstern Are Dead*, Act One, by Tom Stoppard, 1967, Faber and Faber, London—printed here with considerable omissions and slight modifications (not indicated individually) by the authors of the present report)

Contents

1 Introduction

This is the second part of a two-report mini-series focussing on issues in the evaluation of annotations. In this second, more empirically-oriented part, we describe the project background, present results from the practical application of the relevant statistics and, of course, discuss our respective results. The first, more theoretically-oriented part [Stegmann and Lücking, 2005], comprises a summary of the relevant statistical background, an evaluation of some pertaining approaches and a sketch of arguments that may lend themselves to the development of an original statistic. Thus, in fact, the two reports come as a couple, like two sides of a coin. They have been separated in order to allow for a more linear discussion of the general theoretical issues (in the other report) and the setting-specific application issues (in this report), respectively. We believe that the points that are made in this part will be of interest to people involved in empirically-grounded research on task-oriented and/or multi-modal dialogue. Generally, this document is the third one in a series of technical reports authored by linguists of the B3 "DEIKON"[1] project. Taken in conjunction with its sister report, it continues the direction set forth by the first report in the series [Kühnlein and Stegmann, 2003] in its aim of discussing empirical issues with an eye towards the theoretical underpinnings, as well as the practical consequences of the findings obtained.

These publications are complemented by those of our project partners in computer science. Hence, both realms—and the interface between them—are covered: empirical, theoretical and computational aspects of the linguistic integration of speech and deictic gestures, as well as its (re-)synthesis and recognition by means of artificial intelligence methodology in a virtual reality setting [Kühnlein et al., 2003, Rieser, 2004, Kopp and Wachsmuth, 2004, Kranstedt et al., 2002, Kühnlein et al., 2004]. Last but not least, it should be noted that some of the results that will be reported below have been touched upon (among other statistical results) in a workshop contribution [Lücking et al., 2004].

Linguists working on dialogue data have to overcome at least two characteristics of human language in it's primary form: firstly, spoken language is an ephemeral phenomenon which exhales barely nascent, and secondly, natural language is, for the most part at least, not endowed with explicit structural and content markers (e. g. tags that indicate discourse segment boundaries or labels that name the performed dialogue moves).[2] To dispose of the first problem, natural language data are often conserved by using recording techniques, such as audio taping, video filming, or applying systems of manual transcrip-

[1] DEIKON is an acronym for the German project title "DEIxis in KONstruktionsdialogen".

[2] Though it is possible to include signals that carry meta-communicative content such as "Now I start a new discourse segment", such language use can be disregarded not only as being marginal, but also as highly artificial.

tion. This stage of empirical work comes with its own difficulties (potential problems include—but are not limited to—ensuring ecological validity, issues in the protection of the private spheres of the individuals involved, and the ever-present danger of overly theory-ladenness of observation), but we will not be concerned with these matters to a greater degree of detail here. Rather, our main topic is bound up with the second problem that has been mentioned above: Since linguists are often interested in language properties that are covert, respective data can't be merely recorded, they rather have to be "produced". A typical *data-generating source*[3] is the linguist involved, who acts as a *rater* of the bare recorded data. Thus, empirical work in linguistics often builds on the subjective judgements of the researchers themselves. One of the vehicles raters use to furnish their data, i. e. part of the *data-collection method*, is the *annotation scheme* (plus the instructions on how to make use of it). Such a scheme provides a "classification blueprint" which regulates how to classify the phenomena under observation with respect to predefined categories and rules of allocation. In order to indicate that data augmented in this way fulfill the usual scientific requirements, e. g. reproducibility, one has to take steps in order to assure the quality of the scheme. As Carletta [1996, p. 249] puts it from the perspective of dialogue research in computational linguistics:

> Now researchers are beginning to require evidence that people besides the authors themselves can understand and make the judgements underlying the research reliably. This is a reasonable requirement because if researchers can't even show that different people can agree about the judgements on which their research is based, then there is no chance of replicating the research results.

What this report addresses, then, is measuring the reliability of an annotation scheme for multi-modal communication in task-oriented dialogue. Generally speaking, reliability will be construed as the degree of accordance between the output of data-generating sources with respect to the same range of subjects. This formulation is surely very across the board and somewhat naïve, but it would be rather pointless to try to elaborate on it unless some words have been said pertaining to the different aspects of reliability, as well as the different kinds of data generated. This, however, is a task that is accomplished in the sister report, cf. [Stegmann and Lücking, 2005].

This report is organized as follows. Chapter 2 introduces the empirical setting and the core dialogue data. Further, the applied annotation scheme and the data produced thereby are described. Our outcomes are presented and put into perspective by means of interpretation and discussion in chapter 3. Of course, the influence of our findings for present theoretical and empirical work in the project will be underlined. We will finally have some words to say with respect to the future make-up of the scheme in chapter 4.

[3] By using the general terms "data-generating source" and "data-collecting method" we adopt the terminology of [Gwet, 2001].

2 DEIKON in a Nutshell

This chapter serves to introduce some cardinal themes of the research undertaken by linguists in the DEIKON project. Our point will be realized with a focus on both the empirical setting exploited and the theoretical approach chosen (section 2.1), as well as the annotation scheme applied to the data (section 2.2).

2.1 Theoretical Approach and Empirical Setting

Most prominent among the theoretical aims of the DEIKON project is the development of a theoretically adequate account of multi-modal utterances in task-oriented dialogue [Rieser, 2001, Kühnlein et al., 2003, Rieser, 2004]. With respect to such matters, our general theoretical frame can be characterized as being motivated by a certain philosophical perspective [Rieser, 2001, 2004] that we make reference to as a *neo-Peirce-Wittgenstein-Quinean* stance [Hartshorne and Weiss, 1967, Wittgenstein, 1958, Quine, 1960], compare also the semiotic perspective of [Clark, 1996]. This perspective involves that we construe the meaning of "language" in a broad sense, that is, subsuming both the use of gestures and speech. Accordingly, acceptable structures for means of representation have to include entities that accord to the non-verbal elements of such complex signs. Straightforwardly, our current theorizing presumes that manual gestures should be represented on a par with verbal signs at the level of a flattened, modality-neutral communication channel [Kühnlein et al., 2003]. Furthermore, we conceive of multi-modal utterances as being representable by linear successions of symbol-like entities at such a stage. Example (2.1) below (which translates to (2.2)) shows how we conceive of the representation of a simple utterance token in this manner.

(2.1) die rote ↘ Schraube

(2.2) the red ↘ bolt

Here, the "↘" [Rieser, 2001, 2004] represents the deictic gesture's stroke, i. e. the meaningful part of the pointing gesture. It can be placed among the linguistic tokens due to ranking with respect to certain anchoring time stamps, e.g., the start times of the entities involved [Kühnlein and Stegmann, 2003].[1] An abstract representation level as such is necessary as an interface format for linguistic formalisms to operate upon, since those entities that are fed into a parser or do get generated from an appropriate base [Stegmann, 2004] are "structurally innocent" strings of characters in tokenized form,

[1] Pertaining to our example utterance above: the stroke of the deictic gesture which accompanies the uttering follows the onset of the adjective, but precedes the onset of the noun. Deliberately, finer-grained temporal information is not taken into account.

i.e. sequences of words.[2] As has been hinted at above, we assume that gestures make a difference concerning the meaning of multi-modal utterances [Clark, 1996, McNeill, 1992]. This point is also corroborated by our findings in the empirical realm [Lücking et al., 2004]. For example, we found that "gestures save words": utterances that include gestures are shorter than those without them when communicating the same amount of information. We take this as indicating that the pointing gesture carries a significant amount of information, which contributes to the verbal information of the utterance. This point can also be taken to stress the necessity for the development of truly multi-modal grammars, since grammars that are constrained to the verbal elements alone fail to render such facts explicable. Accordingly, one main aspect of our theoretical work consists in the specification of grammars that are capable of handling the interaction between verbal speech and manual pointing, as appropriate with respect to the empirical data [Kühnlein et al., 2003]. We decided to do so on the basis of a constraint-based lexicalist framework of generative grammar that is grounded in the HPSG ("Head-Driven Phrase Structure Grammar") framework [Pollard and Sag, 1987, 1994, Sag and Wasow, 1999]. Current and still ongoing implementations make use of the LKB ("Linguistic Knowledge Building") grammar development environment [Copestake, 2002].

As has been remarked above, our theoretical modelling is constrained by observations made on grounds of empirical studies that have been conducted in a task-oriented setting, cf. [Kühnlein and Stegmann, 2003] for a detailed description of the setting. To give a modestly detailed account of the relevant issues here: the setting of the *pre-test studies*[3] realizes a deviation from the "standard" SFB setting [cf. Corpus]. In contrast to the dynamic standard scenario, it involves the task of merely establishing reference in simple identification tasks. The domain of reference consists of 32 wooden pieces which form the proper components of the "SFB toy airplane"—cf. Figure 2.1. They are referred to collectively as the *baufix domain* here. The toy airplane components are located on a table, where they are distributed in a uniform way and in accordance with a certain pattern (interchangeably: similarity of color *vs.* similarity of form). Compare figure 2.2, which depicts the baufix domain according to a color pattern used. The whole pre-test line consists of seven quasi-experimental runs, where each run involved two subjects, who were assigned the roles of *Instructor* and *Constructor* one after another. The latter role-models can be summarized as follows: *Instructor* is expected to choose an object from the baufix domain and has to convey her choice to *Constructor*, e. g. by producing an appropriate multi-modal utterance involving pointing and speech.[4] *Constructor*'s task,

[2] Of course, this does not mean that the processes involved in parsing and generating do not deal with structures—on the contrary! Those entities that are complementary to our "structurally innocent" ones with regard to the output in parsing (parse tree) and the input in generating (semantic input representation) are essentially structural.

[3] The name "pre-test studies" is due to their quasi-experimental character. These studies were not performed with strict analytical hypothesis testing in mind, but rather as a source of inspiration for hypothesis invention and theorizing.

[4] The subjects received no explicit instruction to use speech as well as pointing, but whatever means of communication would seem appropriate to them. This approach was chosen in order to avoid artificiality effects. However, this strategy resulted in a comparably small basis of usable data, since to our surprise most subjects did not—or only seldom—make use of pointing gestures.

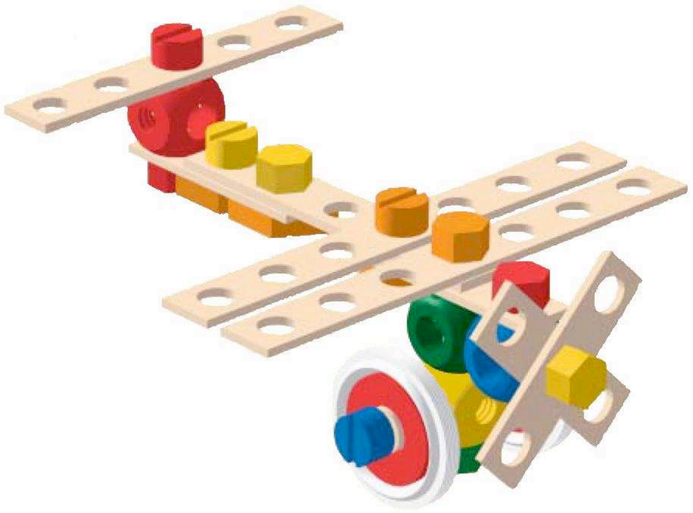


Figure 2.1: The *Lorenz baufix* toy airplane

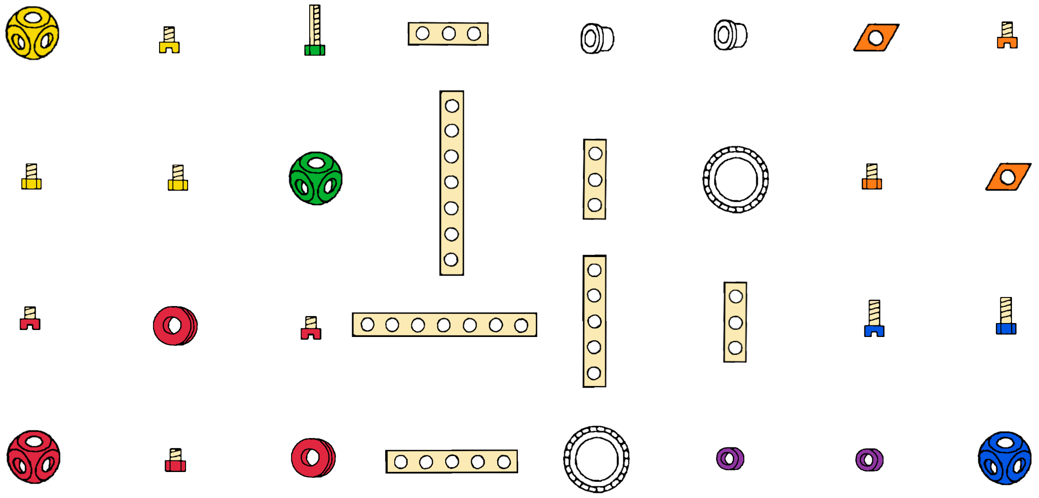


Figure 2.2: The *baufix* domain

on the other hand, consists in trying to resolve the meaning intended by *Instructor* and in delivering appropriate feedback, i. e. concerning which object he assumes *Instructor* intended to refer to. The pre-tests were recorded from different camera perspectives (overall scene view and *Instructor*'s perspective) using digital video equipment. Figure 2.3 gives an impression of the elicited data.

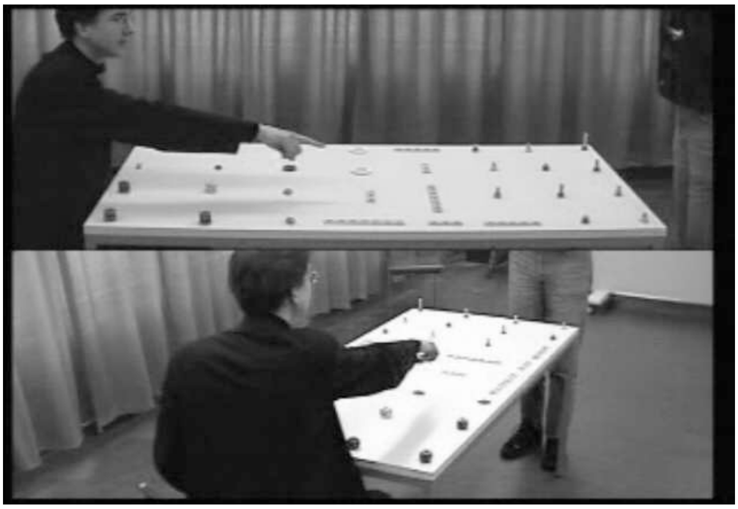


Figure 2.3: Sample still from pre-test 5

One notable result of our analysis of the available data is that the timing relations among gesture and speech are more complex than what was expected in the light of the pertinent literature: cf. the "canonical view" of [McNeill, 1992] and compare our detailed findings that have been reported in [Kühnlein and Stegmann, 2003] and [Lücking et al., 2004]. This means *inter alia* that we have reason to believe that an appropriate grammar for multi-modal expressions should not only be able to generate and parse expressions of the form in example (2.1) above—rather, the following configurations should also be rendered as well-formed and meaningful:

(2.3) ↘ die rote Schraube

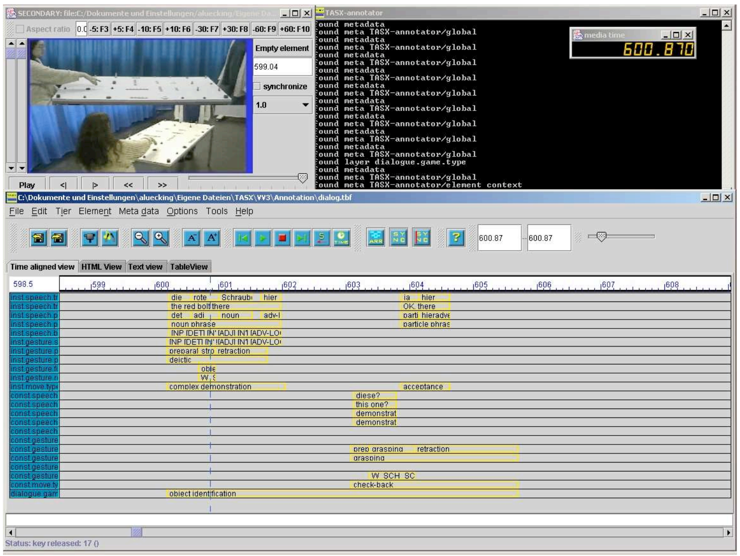
(2.4) die ↘ rote Schraube

(2.5) die rote Schraube ↘

Now, configurations such as (2.3) and (2.5) pose a serious threat to our linguistic theorizing, since, of course, it is desirable to provide a uniform and parsimonious analysis concerning the distribution of the "↘". However, when the "↘" precedes or follows a saturated nominal phrase, while being located in between its constituents for the other configurations, it becomes at least hard to see how the meaning of the head noun can be linked to the meaning of the stroke.

2.2 The Annotation Scheme at Stake

In order to facilitate the analysis and evaluation of the collected data, we had to start annotating them. A decision was made to use the TASX-Annotator software[5] [Milde and Gut, 2001], which is available under the auspices of the *GNU General Public License* and allows for the pursuit of an XML-based bottom-up approach toward the annotation of digital video data [Kühnlein and Stegmann, 2003, ch. 3]. Figure 2.4 below shows a screen-shot from a TASX annotation session. The respective annotation tiers, which are exemplified by means of appropriate event labels according to the example utterance from the preceding section, include:



```
found metadata
found meta TASX-annotator/global
found metadata
found meta TASX-annotator/global
found metadata
found meta TASX-annotator/global
found metadata
found meta TASX-annotator/global
found metadata
found meta TASX-annotator/global
found metadata
found meta TASX-annotator/global
found metadata
found meta TASX-annotator/global
found metadata
found meta TASX-annotator/global
found layer dialogue.game.type
found metadata
found meta TASX-annotator/global
found meta TASX-annotator/element context
```

Figure 2.4: Annotation with TASX-Annotator

speech.transcription die, rote, Schraube, ...

speech.translation the, red, bolt, ...

speech.pos det, adj, noun, ...

speech.phrase noun phrase, ...

gesture.phase preparation, stroke, retraction, ...

gesture.phrase deictic, ...

gesture.function object pointing, ...

gesture.reference W_SCH_SCHRAUBE-rot-0, ...

[5] Homepage: http://tasxforce.lili.uni-bielefeld.de/

move.type complex demonstration, ...

game.type object identification game, ...

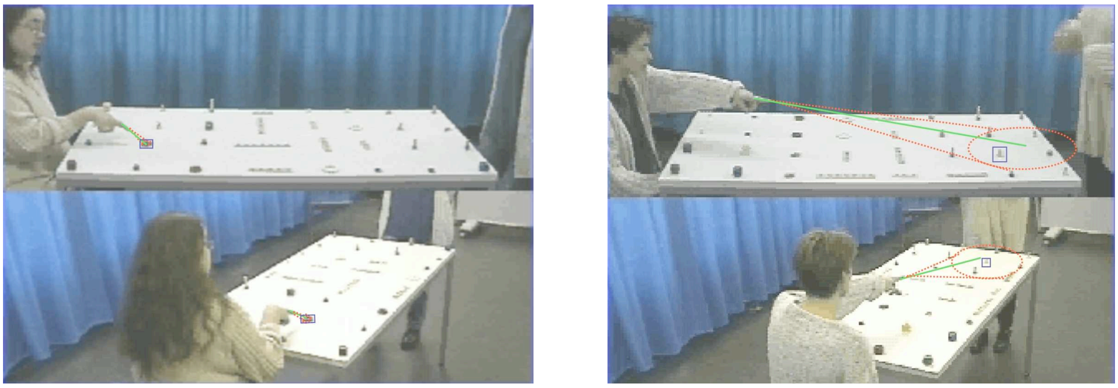
Furthermore, all annotation tiers carry an `inst.`- or `const.`-prefix (except for `game.type` since it applies to both). The prefix indicates whether the respective property is ascribed to *Instructor* or *Constructor*. This naming convention was chosen with an eye toward XSLT-internal regular expression capabilities for means of efficient post-processing of the annotated data.

We shall delve into the internals of the scheme now. There are three main classes of observable phenomena, aspects of which have been annotated: these are, of course, gesture and speech, and furthermore the structure of the dialogue games played in terms of the employed dialogue moves.

Foundational aspects of speech include a transcription of the verbal tokens at `speech.transcription`, English translations of the latter elements, mainly for presentational purposes, at `speech.translation`, classification of word tokens with respect to part-of-speech categories at `speech.pos`, and, finally, the `speech.phrase` tier. There is not much to say about those speech-related annotation tiers, except for the suffix "`phrase`", the latter being somewhat misleading, since the main purpose of this tier does not consist in marking and naming distinct linguistic projection levels, i. e. layers of syntactic analysis. Rather, this tier's function consists in providing a grip at the time boundaries of the verbal wholes of complex utterances, which is important for algorithms underlying XSLT-based processing and analysis means. However, the marked information can be expected to coincide with a certain level of phrasal projection, at least as far as grammatical expressions are concerned.

We now turn to the gesture-related tiers. Following [McNeill, 1992], at the `gesture.phase` tier we distinguish among three main phases most gestures do comprise, namely: preparation, stroke, and retraction. The stroke is the "meaningful", that is, obligatory part of every gesture, whereas preparation and retraction can be considered as anatomical necessities in carrying out the gestural act. This level of classification comes with its own problems, which are related to issues in the applicability of the underlying McNeillian scheme—for a detailed account of such problems with respect to deictic gestures compare [Kühnlein and Stegmann, 2003]. On a higher hierarchical level, a cluster of several gesture phases (i. e. the whole of a gesture in the usual sense of the word) forms a gesture phrase, which finds classification according to its type at the corresponding level, i.e. `gesture.phrase`. In addition, we label each stroke occurrence according to its resolution quality at the `gesture.function` tier. Here, we distinguish between *region pointing* and *object pointing* [Rieser, 2001]. A gesture qualifies for the function of object pointing if it succeeds in establishing the connection to a single object in its own right, i. e. disregarding the information that was contributed by the verbal elements of the utterance. In opposition, gestures that fail to do so will fulfill region function—they merely carry information regarding a certain set of candidate referents of the utterance. The candidates become prominent in honor of lying in that region that is made salient by the pointing.

To be more specific about the point of gesture functions: one of our guiding assumptions is that the meaning of a deictic gesture can be modelled by determining the intersection of the deictic *pointing cone* with the surface of the *pointing table* (compare Figure 2.5 for an illustration).[6] Operationalized as such, object pointing corresponds to one and only one object lying within the intersection, while region pointing complies to several objects lying therein. Now, taking this for granted, what function a gesture fulfills relies (besides the goodness of the pointing, of course) on the distance of the target object from the agent, as well as on the density of the objects in the intersection region (cf. the respective remarks in section 3.3 below and also [Lücking et al., 2004]). As should be obvious, in our setting the latter is related to the former—to state the respective rule of thumb: "the longer the distance of the object from the agent, the higher the density in the target region, the worse, *ceterus paribus*, the resolution of the pointing".



(a) Object pointing

(b) Region pointing

Figure 2.5: Depiction of the pointing cone

Finally, we have the `gesture.reference` tier that fulfills the purpose of naming the intended target of the complex reference act in an unambiguous way (the terms that function as values stem from the vocabulary of a specialized knowledge representation language, cf. [Jung, 1996]). Note, that for cases of region pointing the intended referent may only be recognizable by looking at the outcome of the dialogue game wherein the utterance is embedded[7]. Specifications at this annotation tier are necessary for bringing about the XSLT-based transformation from our annotation format to the kind of representation scheme that is used by our project partners in computer science for

[6] Results driven by certain domain-specific hypotheses and idealizations indicate that the effective resolution of the pointing finger lies somewhere between six and twelve degrees [Kühnlein and Stegmann, 2003]. However, this estimate seems to be biased with respect to *Instructor*'s behavior. Further, the results of our reliability study suggest that the estimations applied by the raters show considerable variability.

[7] This presupposes, of course, that the instructor does not change her mind midways with respect to the originally intended referent of the utterance. However, we think that this is a reasonable hypothesis for the vast majority of cases.

the sake of simulating respective pointing gestures—hence the placement of this tier in the hierarchy below `gesture`.

The classification of dialogue move types on the `move.type` tier was done in a preliminary way, i. e. no sober worked-out game-coding scheme has been applied. The basic purpose of the labelling of moves was (a) to illuminate our way of thinking about the structure of the object identification games observed, and (b) in order to help us make up our minds concerning the option of adapting an already established proposal for dialogue annotation, e. g. the DAMSL [Core and Allen, 1997] or the HCRC coding scheme [Carletta et al., 1997]. The whole set of move-type labels used for the present purposes consists of the following nine plus one classifiers, each given with a few notes and an exemplification from our corpus here. English translations are added for convenience.

clarification The subject asks a clarifying question.

> *Welche?*
> Which one?

repair-of-clarification The subject puts a former clarification in more concrete terms.

> *[Welche?] Mit den vielen oder mit den wenigen Löchern?*
> [Which one?] With the many or with the few holes?

assertion The subject states something.

> *Du meinst den gelben Würfel.*
> You mean the yellow cube.

complex-demonstration The subject performs a multi-modal reference act, that is, he or she uses gesture plus speech to refer to an object.

> *Die gelbe ↘ Schraube.*
> The yellow ↘ bolt.

demonstration The subject refers to an object *via* a definite description, that is, he or she uses solely speech, no gesture.

> *Die gelbe Schraube.*
> The yellow bolt.

check-back The subject queries whether he or she identified the right object.

> *Diese ↘?*
> This ↘ one?

act/check-back The query is performed without speech, by lifting an object up or pointing at it.

↘ (same gesture in English)

acceptance One subject affirms the other subject's choice of an object.

Ja. – Richtig.
O.K. – Right.

repair The subject emends his or her own former utterance.

[Die grüne Schraube.] Nein, die rote.
[The green bolt.] No, the red one.

? none of the categories above (unclassifiable)

Note that the ten move-type labels do not distribute equally over Inst's and Const's moves. The *Instructors*' moves comprise: clarification, repair-of-clarification, assertion, check-back, act/check-back and acceptance. In opposition, the constructors' dialogue contributions are exhausted by: assertion, complex-demonstration, demonstration, acceptance, and repair. In addition, both instructor and constructor produced utterances that had to be rated as unclassifiable by at least one of the raters. We will come to criticize the current makeup of the dialogue scheme in conjunction with the results on reliability in the final chapter of this report.

The following example dialogue, taken from Figure 2.4, will convey a better idea of how a typical object identification game expands over time:

Inst: *Die rote ↘ Schraube hier.*
The red ↘ bolt here.

Cnst: *Diese ↘?*
This one ↘?

Inst: *Ja hier.*
Yes here.

The *Instructor* starts the game with a multi-modal reference act concerning a certain object, that is, he uses a *complex demonstration*. *Constructor*, then, performs a *check-back* by lifting an object and asking whether it is the right one. Since the object was correctly identified, the instructor can terminate the game with an *acceptance* move.

The annotation task on all the tiers described involves the determination of the start and end points of respective events pertaining to the common time-line of the video recordings. Thus, our annotation scheme is designed to let us achieve insights into temporal relationships of interest.[8] Besides their purpose to carry annotation content, the annotation tiers `move.type` and `game.type`, like the afore mentioned `speech.phrase` and `gesture.reference` levels, are exploited in order to facilitate XSLT-based processing means. Since some moves comprise speech as well as gesture, the respective `move.type` elements "unify" those communicative entities (gestural and spoken alike) that belong to the same conversational move. In a similar fashion, the `game.type` elements "delimit" those moves that constitute a single complete dialogue game. The algorithms that underlie the processing of our XML-annotated data are tuned to the available information on the different tiers. There are two main reasons for XSLT-based transformation: firstly, we have to process our data in order to be able to analyze them, that is, the relevant information has to be extracted and transformed before we can make use of it. For example, the calculation of the statistical results below depends on transformations that extract the relevant temporal and categorical data and lists them in an appropriate format that can be fed into the R statistical programming environment. Secondly, it is desirable to be able to transform our XML dialect (*tasx* format) to another XML variant, i.e. the MURML language [Kranstedt et al., 2002], which is used by our DEIKON project partners in computer science. Transformed as such, original data from the empirical studies can be used to specify the communicative behavior of an embodied conversational agent in a virtual reality setting [Kopp and Wachsmuth, 2004]. Further, such input can be utilized for means of "benchmarking" the behavior repertoire of the virtual agent and vice versa. In this respect, our annotations have been shaped so that they adhere to informational necessities in order to realize the desired transformations and also to meet the MURML specification. In turn, findings from the empirical studies adviced certain design revisions concerning parameters dealing with space and time in the virtual reality setting [Kühnlein and Stegmann, 2003, Kühnlein et al., 2004].

[8] Figuring most prominently among them are gestural anticipation and synchrony, cf. the explanation in [McNeill, 1992] and the respective results reported in [Kühnlein and Stegmann, 2003] and [Lücking et al., 2004].

3 Statistical Results and their Discussion

This pre-final chapter of our report is organized as follows: in the first two sections, we discuss the inter-rater agreement results for data set on nominal scales (section 3.1) and the correlation results for data set on magnitude scales (section 3.2). Then, in the third subsection we interpret the obtained results in perspective and comment on their impact for present and future work in the project (section 3.3). The presentation of our results below builds on the background in reliability statistics that is provided by the sister report [Stegmann and Lücking, 2005, Chapter 2].

3.1 Statistical Results for Inter-rater Agreement

We will start with the perhaps most controversial result that has been achieved: it concerns the presumed gesture function, i. e. *object vs region pointing*, compare the remarks in section 2.2 above. Remember, deictic gestures have to be rated as fulfilling object function when they are "sharp" enough to single out one and only one object, otherwise they are labelled as fulfilling region function. Table 3.1 summarizes the compared ratings concerning the aspect of *reproducibility*, i. e. *inter*-observer reliability, in the form of a contingency table for the 57 gesture tokens rated by our two experts.[1]

Table 3.1: Contingency Table for "Object- *vs.* Region Pointing"

Rater 1	Rater 2		Total
	object	region	
object	15	14	29
region	2	26	28
Total	17	40	57

Table 3.2: Contingency Table for the Stability of "Object- *vs.* Region Pointing" for Rater 1

Rater 1.old	Rater 1.new		Total
	object	region	
object	12	6	18
region	2	5	7
Total	14	11	25

[1] Of course, all the contingency tables that will be listed in this chapter satisfy the scheme laid out in [Stegmann and Lücking, 2005, subsec. 2.2.1]. Hence, our actual results can be verified easily by substituting the respective terms in the formula templates.

Table 3.3: Contingency Table for the Stability of "Object- *vs.* Region Pointing" for Rater 2

Rater 2.old	Rater 2.new		Total
	object	region	
object	10	1	11
region	4	10	14
Total	14	11	25

Our calculations for the reproducibility of "object *vs* region pointing" result in an AC_1 value of 0.48. Of course, with chance-corrected values close to the middle between 0.0 and 1.0, such a result is likely to be better than what can be expected on grounds of chance alone, if only a reasonable number of tokens have been rated. This suspicion is corraborated by the result of a significance test against the corresponding null hypothesis (H_0: "$AC_1 = 0.0$") that comes out as highly significant ($\alpha = 0.01$, $n = 57$).[2] Hence, the alternative hypothesis H_1 (H_1: "$AC_1 > 0.0$") holds with probability $1 - \alpha$ which equals 0.99 here. So it seems safe to project from our actual ratings in such a way that gesture function rating will be performed on a niveau that is better than chance alone. Straightforwardly, however, not very much is achieved by securing that level (we have already discussed this point in [Stegmann and Lücking, 2005, subsec. 2.2.2]). Our respective results here and below are stated rather for convenience and in order to adhere to the usual scientific practice. When we constrain our interpretation efforts to the closed domain of cases that have been rated by our raters, however, we see that this result is not too good. It indicates that less than half of the actual cases have been reliably agreed upon (i. e., after the chance correction has been performed). Intuitively, this seems to mark a rather weak result that we should wish to improve upon in future annotations.

Further results concerning object- and region function pertain to a second aspect of reliability, compare our discussion in [Stegmann and Lücking, 2005, sec. 2.1]: that is the *stability* dimension, i.e., the *intra*-observer perspective. Therefore, both experts had to re-rate pertinent tokens concerning gesture function. These new ratings were then compared to the old ratings of the same tokens for each rater, compare the corresponding Table 3.2 for Rater 1 and Table 3.3 for Rater 2. What we get is an AC_1 of 0.41 for Rater 1 and an AC_1 of 0.60 for Rater 2. Both results, again, project to be better than what can be expected on grounds of chance alone, as is shown by the result of the significance test for Rater 1 ($\alpha = 0.05$, $n = 25$) and Rater 2 ($\alpha = 0.01$, $n = 25$). This notwithstanding, in the light of a direct interpretation of the absolute values for our closed domain, better results for robustness over time would surely be desirable—especially with regard to those achieved for Rater 1.

[2] The α value gives the upper bound probability for committing an error of the first kind, i. e., rejection of the null hypothesis when it is, in fact, true. Furthermore, of course, it is preferrable to minimize that probability. Hence, the following marks a sound slogan: the lesser the α, the more persuasive the result.

Table 3.4: Contingency Table for Instructor Move Types

Rater 1	Rater 2						
	acceptance	assertion	complex dem.	demonstration	repair	?	Total
acceptance	34	–	1	–	–	–	35
assertion	1	–	–	–	1	–	2
complex dem.	–	–	40	–	–	–	40
demonstration	1	–	–	9	–	–	10
repair	1	–	2	–	1	–	4
?	–	1	–	–	–	–	1
Total	37	1	43	9	2	0	92

Table 3.5: Contingency Table for Constructor Move Types

Rater 1	Rater 2							
	acceptance	act/ckeck-back	assertion	check-back	clarification	repair of clar.	?	Total
acceptance	–	–	–	–	–	–	–	0
act/check-back	3	16	–	–	–	–	–	19
assertion	–	–	–	–	–	–	–	0
check-back	1	2	2	33	1	1	–	40
clarification	–	–	–	–	4	1	–	5
repair of clar.	–	–	–	–	–	–	–	0
?	1	–	–	–	–	–	–	1
Total	5	18	2	33	5	2	0	65

Move type classification at the level of the *dialogue acts*, cf. the details explained in section 2.2 of this report, proved to be more reliable with regard to the aspect of *reproducibility* (*inter*-observer perspective, again). Firstly, the AC_1 for Instructor's move types comes close to a perfect result with a value of 0.90. Secondly, the result for rating the more variable Constructor's move types equals 0.795, which is still a very good value. The cross classifications for the raters' judgements are given in tables 3.4 and 3.5. With respect to significance tests both results come out as very highly significant against the chosen null hypothesis of no agreement apart from chance agreement (Instructor's moves: $\alpha = 0.001, n = 92$; Constructor's moves: $\alpha = 0.001, n = 65$).

But would the raters be able to repeat their classification of move types in a consistent way? About half a year after the first annotation session, both had to re-annotate the videos with regard to the same dialogue coding scheme. AC_1 was used again in order to compare the old ratings with the new ones: thereby a value for the *intra*-rater reliability or *stability* was calculated. To particularize, Rater 1's value for the stability of rating Instructor's moves comes to an AC_1 of 0.89, for Constructor's moves it amounts to 0.75. Concerning the results for Rater 2, with an AC_1 of 0.81 for the classification of Instructor's moves and a value of 0.72 pertaining to the rating of Constructor's moves, the agreement coefficients obtained for Rater 2 are located on a similar good rank as those for Rater 1. Of course, all the mentioned values are very highly significant ($\alpha = 0.001$, $n = 92$ for Instructor's moves; $\alpha = 0.001$, $n = 65$ for Constructor's moves, respectively). More detailed information can be retrieved from inspecting the numbers displayed in the tables 3.6 to 3.9.

Finally, we leave the topic of stability in order to investigate *reproducibility* once more, i. e. the *inter*-rater reliability for the application of the dialogue coding scheme. This time our calculations are performed against the background of the re-ratings and we compare just the new ratings of Rater 1 and Rater 2. Here, the annotation of Instructor's moves comes to an outstanding result of 0.94, while comparing the ratings of Constructor's moves results in a slightly worse AC_1 value of 0.75, as compared to the result obtained for the original ratings. Of course, both are very highly significant against the null hypothesis (Instructor's moves: $\alpha = 0.001$, $n = 92$; Constructor's moves: $\alpha = 0.001$, $n = 65$). When we regard our results on stability and reproducibility for dialogue moves in conjunction and on direct terms, we observe that the results for the rating of Instructor's moves are very good and also superior to those for the rating of Constructor's move types. This notwithstanding the latter are still rather good. Furthermore, it could be shown that both types of results are reasonably robust over time, although there are slight advantages for the rating of Instructor's moves again.

Although timing phenomena have been introduced as being rooted on a magnitude scale—compare our remarks in [Stegmann and Lücking, 2005, sec. 2.3] and the pertinent results in the next section—the strict linearity of the time scale can be broken up in order to take a categorical perspective. This flanking approach was chosen, since our linguistic apparatus is set up in a way such that even minor deviations in the time-line based ratings lead to different symbolic input sequences for parsing to operate upon. Since some sequences are rather problematic, this is an issue of some concern to us, compare our remarks in our sketch of the theoretical approach in section 2.1 above. Now, when the *position of the gestural stroke* is classified into the nominal categories "first", "(somewhere in) between", and "final" pertaining to the insertion of the stroke among the linguistic tokens of the utterance (construed along the exact timing of the respective entities), we obtain a categorical representation of the problem such that the calculation of AC_1 is possible. The transformed data are summarized in table 3.10. What we get is an AC_1 of 0.73 for *reproducibility*, which is not as good as the respective correlation value for the magnitude scale level that will be reported in the next section.

Table 3.6: Contingency Table for Stability Classification of Instructor Move Types by Rater 1

Rater 1 inst.new	Rater 1 inst.old						
	?	acceptance	assertion	complex dem	demonstration	repair	Total
?	–	–	–	–	–	–	0
acceptance	–	32	–	–	1	1	34
assertion	1	3	1	–	–	–	5
complex dem	–	–	–	40	–	1	41
demonstration	–	–	–	–	9	–	9
repair	–	–	1	–	–	2	3
Total	1	35	2	40	10	4	92

Table 3.7: Contingency Table for Stability Classification of Constructor Move Types by Rater 1

Rater 1 const.new	Rater 1 const.old					
	?	act/check-back	assertion	check-back	clarification	Total
?	–	–	–	–	–	0
act/check-back	–	17	–	3	–	20
assertion	–	–	–	–	1	1
check-back	–	1	–	31	2	34
clarification	1	1	–	6	2	10
Total	1	19	–	40	5	65

However, it has to be taken into account that the latter is a statistics that is not corrected with respect to chance. And anyway, this result seems to be good enough in order to strengthen our confidence in the structural implications of time-line based ratings to some reasonable degree. Of course, this result is very highly significant against our null hypothesis ($\alpha = 0.001$, $n = 25$).

To summarize on our results, we will rank the values according to the absolute AC_1 values obtained for our closed domain here. We get the following picture: the least good results are those for the reproducibility (0.48) and the stability (0.41 and 0.60) of the gesture function coding. Following up, the reproducibility result for stroke insertion

Table 3.8: Contingency Table for Stability Classification of Instructor Move Types by Rater 2

Rater 2 inst.new	Rater 2 inst.old							
	acceptance	assertion	complex dem	demonstration	repair	clarification	?	Total
acceptance	35	–	1	–	–	–	–	36
assertion	1	–	–	–	–	–	–	1
complex dem	–	–	41	–	–	–	–	41
demonstration	–	–	–	9	–	–	–	9
repair	–	–	1	–	2	–	–	3
clarification	–	1	–	–	–	–	–	1
?	1	–	–	–	–	–	–	1
Total	37	1	43	9	2	0	0	92

Table 3.9: Contingency Table for Stability Classification of Constructor Move Types by Rater 2

Rater 2 const.new	Rater 2 const.old						
	acceptance	act/check-back	assertion	check-back	clarification	repair of clar.	Total
acceptance	3	–	–	–	–	–	3
act/check-back	–	14	–	2	1	2	19
assertion	1	–	1	–	1	–	3
check-back	–	3	1	29	1	–	34
clarification	1	–	–	2	1	–	4
repair of clar.	–	1	–	–	1	–	2
Total	5	18	2	33	5	2	65

(0.73) as well as those for the stability and reproducibility of constructor's dialogue moves (0.72, 0.75, and 0.795) seem to be of fairly good quality. Finally, the results for the instructor's dialogue moves are best (0.81, 0.90, and 0.94) with the values for the reproducibility aspect being truly impressive.

Conforming with our suggestions in [Stegmann and Lücking, 2005], we have commented on our actual results on grounds of three lines of reasoning here: informed intuitions, internal ranking and the usual significance tests. As we have argued before,

Table 3.10: Contingency Table for "Position of $\searrow$ in Surface String"

Rater 1	Rater 2			Total
	first	between	final	
first	1	1	–	2
between	–	13	–	13
final	–	4	6	10
Total	1	18	6	25

we think that there can be no non-arbitrary standards for judgements concerning the quality of reliability results. Now, interestingly enough, our interpretations seem to fit rather well to what we would have got from the application of Krippendorff's scale, compare the description in [Stegmann and Lücking, 2005, subsec. 2.2.2]. For better or worse, the application of that scale seems to mark a *de facto* standard, at least among "kappa" practitioners in computational linguistics, cf. [DiEugenio and Glass, 2004]. Also, Krippendorff's scale is a very strict one and we said that AC_1 results should be interpreted in a rather strict way, since they tend to be better than those for the other statistics, cf. [Stegmann and Lücking, 2005, subsec. 2.2.2]. This notwithstanding, we still think that it would be a bad idea to apply such a scale blindly, since the borders drawn by such scales are hard ones, e.g., for Krippendorff's scale the "critical values" are 0.67 and 0.80. We think that such delimiting values must not be seen analogous to, say, the critical values of test statistics. In our opinion, no hard categorial distinction should be made on grounds of neighboring results alone, e.g., 0.66 *vs.* 0.68 or 0.79 *vs.* 0.81, as is implied by Krippendorff's perspective. After all, the latter values are just a tiny bit better than the former ones and the details of the categorial disctinctions are clearly arbitrary. This notwithstanding, we think that the general tendency underlying Krippendorff's scale, i.e., to demand for results around 0.66 in order to count as "good" and ones around 0.80 to qualify as "very good" is aiming in the right direction, since such seem to be intuitively appealing niveaus. Indeed, agreeing on more than two-thirds of the cases (or, for the better level: four-fifths) seems to be a good result (or: a very good one, respectively), while agreeing on less than half of the cases does not seem to be too reliable. Our comments above on the actual results adhere to this perspective. We will speculate about how some of our results, good and less good ones alike, can be explained and have something to say about possible consequences in section 3.3 below.

3.2 Statistical Results for Correlation

Using Bravais' and Pearson's product-moment correlation coefficient r_{xy} we have calculated the agreement between the rater's scales for the boundaries of word and gesture phases. We have supposed a confidence level of 0.99, that is a two-tailed standard error $\alpha = 0.01$.

Table 3.11: Results for boundaries of gesture phases

	preparation		stroke		retraction	
	start	end	start	end	start	end
d	35117.97	35092.56	35092.56	35056.78	35056.78	35076.45
r	0.9999999	0.9999999	0.9999999	0.9999998	0.9999998	0.9999976
r^2	0.9999998	0.9999998	0.9999998	0.9999996	0.9999996	0.9999952
N	25 for all cases					

Table 3.12: Results for word boundaries

	start	end
d	34198.83	34206.14
r	0.9999999	0.9999999
r^2	0.9999998	0.9999998
N	108 words	

With values of r_{xy} very near to 1, we found that the ratings of boundaries are highly significant under the choosen condition of $\alpha = 0.01$.

A comment has to be made regarding reentrant values in the results for gesture phase boundaries. Since a gesture is a continuous act, the classification of the three gesture phases splits the gestural movement in a certain point in time, such that the end of the preceding part is identical to the start of the next one. Nonetheless, for the sake of completeness and lucidity we listed all resultant values.

In order to say something more detailed about the relationship between the ratings of boundaries, we computed the linear regression of Y onto X, where Y stands for the scores generated by rater 1 and X for the scores produced by rater 2. Calculating regression functions adheres to the general formula $y = a \cdot x + b$, where a defines the slope and b is the intercept term. The outcomes for the gesture phases are given in table 3.13. As might be expected in light of the very high values for r, there are no “runaway formulas”, i. e., the slopes approximate the value of 1 very closely, and the intercept terms tend to be zero. With values of roughly -0.28 and 0.67, only the intercepts of the

Table 3.13: Regression function for gesture phase boundaries

preparation	start	$Y = 0.9999465 \cdot X - 0.07956419$
	end	$Y = 0.9999826 \cdot X + 0.1149759$
stroke	start	$Y = 0.9999826 \cdot X + 0.1149759$
	end	$Y = 1.000388 \cdot X - 0.2822953$
retraction	start	$Y = 1.000388 \cdot X - 0.2822953$
	end	$Y = 0.9992468 \cdot X + 0.6666645$

start and the end of the retraction protrude. This indicates that rater 2 usually marks the termination of the respective gesture phase a bit earlier than rater 1, but estimates their beginning a split second later.

The regression functions for the classifications of words' beginning and ending are given in table 3.14. Being opposite to the start of the retraction phase, rater 2 tends to hear the words' initial sounds slightly earlier than rater 1.

Table 3.14: Regression function for word boundaries

word	start	$Y = 0.998845 \cdot X + 0.376646$
	end	$Y = 1.000011 \cdot X - 0.04468603$

3.3 Discussion of our Results

This section elaborates on the results presented in the sections above and tries to put them into perspective. To summarize on our findings: What could be observed very well, is a near perfect level of correlation pertaining to the ratings of start- and end times of time-stretched entities, i. e. words and gesture phases. However, when considered from a more fine-grained perspective, one that looks at the consequences of minor disagreements concerning timing—deriving the surface position of $\searrow$ with respect to it's linguistics affiliates—slight deviations have to be noted in the results for inter-rater reliability. However, the value achieved is still a good one, albeit not very good. Matters are worse with regard to the results concerning presumed gesture function that come out as rather weak. In opposition, results for dialogue act classification are generally good, with those for Instructor's move types being very good. We will discuss some of these findings in greater detail now.

Firstly, there is not much to say about the ratings of the beginnings of words and gesture phases. The respective correlation values are too good to worry much about. Just two observations could be held, which ensue from the calculation of regression functions (compare subsection 3.2). If one can speak of difficulties at all, delimiting the retraction phase seems to be more difficult than delimiting the other gesture phases. Especially the determination of the end of the retraction seems to be comparably tough (the intercept term is about 0.67). The retraction phase ends when a rest position is reached. But usually the subjects don't let their arms rest, they rather remain in motion. So the intricate question the raters had to answer concerns when exactly the gestural movement ends and the "rest movement" begins. Apparently, in deciding on this matter, the raters have come to slightly different results.

The second "difficulty" concerns the assessment of words' startings. Having no problems to determine the end of the words' final sounds, there is some disagreement between the raters on when to mark their onset. We are not sure whether this asymmetry is a systematic one—probably not. The single phones that constitute an utterance are not clearly separated; they are merged due to various co-articulatory effects. Perceiving and identifying phonemes is governed by acoustic properties such as formant frequency, voice onset time, intensity, friction, and voicedness. The noticeable features of a single phoneme, in turn, are highly context-sensitive, i. e. they are affected by their neighboring sounds. Against the background of phonetics, speaking not too seriously, it is almost astonishing that raters who have to delimit word boundaries are able to agree at all! Since detailed phonetical investigations would lead us too far from our main concerns, we refrain from further considerations in this regard and remain pleased with our respective results.

Matters come off worse when turning to the classification of gesture function. A common cause for low agreement in classification tasks, as is the case with the distinction between object and region pointing, is that type-ii data give leeway to idiosyncratic judgements. Not all subjects represent the characteristics that match the definition of the response category in the annotation scheme in a perfect way. In such a "rating under uncertainty", different raters may use divergent heuristics in assigning the data to the categories. To bridge to the DEIKON studies, our raters seem to apply observed features to defined features in variable and varying ways to pass their sentences about gesture function. But the disappointing agreement value ($AC_1 = 0.48$) bears the opportunity to clarify the criterions a pointing device and the pointing situation must exhibit in order to be classified as region or object pointing. Since there is a mild level of agreement, a closer look at the gesture function ratings might be promising. The detailed outcomes are given in Table 3.15. A striking quantitative pattern is that there is perfect agreement concerning gestures pointing into columns 1 and 2 and into columns 7 and 8, except for one outlier in column 8. Furthermore, for the first two columns nearly all gestures are classified as object pointing, whereas in the last two columns there is a clear dominance of region pointing. Agreement diminishes for columns 3 to 6. Clearly, the distance of the objects pointed at seems to be a distinguishing feature for gesture function. And in addition, since object pointing is conceived of as being "sharper" than its region counterpart, gestures seem to be "distance-sensitive" in the sense that gestures get fuzzy at distance. Note that these properties are explainable by the presumed model of a deictic "pointing cone". The gestural "fuzziness" may also be influenced by properties of the pointing domain like matters of salience and densitiy, cf. the remarks given in [Lücking and Rieser, 2004]. In addition, the exact degree of fuzziness seems to be perceived differently by our two raters. We hypothesize that it would also be judged differently by Instructor and Constructor. Therefore, it seems to be appropriate to presume that the pointing cone will be construed differently relative to the various perspectives [Stegmann, 2004]. Furthermore, an aggravation for a reliable application of the proposed categories

Table 3.15: Gesture function classification

	Column	1	2	3	4	5	6	7	8
Rater 1	Object pointing	2	4	8	6	7	1	0	0
	Region pointing	0	1	2	1	3	9	7	5
Rater 2	Object pointing	2	4	6	2	2	0	0	1
	Region pointing	0	1	4	5	8	10	7	4

to a reasonable number of cases is surely the quality of the digital video data. The three-dimensional real life dialogues are reduced onto a two-dimensional representation: with respect to the assessment of spatial qualities this seems to be a serious drawback. But this notwithstanding, we still have to show that the distinction can be applied to the problematic cases in principle. Remember that we have to do with theory-internally developed categories here and that we have construed reliability as a necessary, albeit insufficient condition with respect to validity.[3] Therefore, the presently weak results pose a threat to the validity of certain aspects of the theoretical framework. Cooperative, B3 enfolding work on the topic of the "pointing cone" that will rely on tracking studies and simulation of empirical pointing data, will hopefully deliver further insights concerning this difficult topic.

Further qualifications have to be made concerning the annotation scheme for dialogue acts. Although the blank results for reliability are very convincing, it has to be noticed that the pattern of dialogue moves that were used by the subjects for means of solving the task we set for them looks, by and large, rather uniform (cf. the description of move type classification and the structure of object identification games in chapter 2). The overwhelming majority of Instructors' moves consists of complex demonstrations, demonstrations, and acceptances; with Constructors' moves, check-back and act/check-back have the lion's share—compare the rating tables given as Table 3.4 and Table 3.5. Accordingly, the set of the regularly employed move type labels is comparatively small. In must be taken into account that the lesser the number of rating categories, the lesser the chance of disagreement between the raters[4]. So few move type categories foster the agreement coefficient to take on a higher value, especially with the chance term being estimated as comparably small due to the many open categories (most of them are used very seldom). Also, many games follow a stereotypical pattern, that does not leave much room for variation and, hence, for disagreement among the raters. Taking all of this into account together with the preliminary and explorative character of our current move type scheme (cf. the respective notes in chapter 2), there seem to be two options for improvement. As the first alternative, we could adopt an already well-established dialogue coding scheme to fit to our setting. This move would have the advantage of entailing that our annotations would be easier accessible to other researchers. However, it has to be noticed that the reliability results for the best known candidates, namely the HCRC- and the DAMSL coding scheme, are not better or even worse than the results for our scheme. With DAMSL, *kappa* values calculated for the various move types applied range from 0.14 to 0.77 [Core and Allen, 1997]; the overall result for HCRC amounts to 0.83 [Carletta et al., 1997]. But then again, it must also be taken into account that the HCRC and the DAMSL scheme might come to achieve better results in a more restricted dialogue setting as ours. Furthermore, when comparing such results it has to be remembered that the usage of *kappa* may often result in comparably lower overall values, due to the over-estimation of the error term incorporated. The alternative to

[3] Indeed, this is a rather verificationist outlook, but an adequate one, we think.

[4] The relation between the number of categories and the chance for disagreement can be strikingly pointed up when considering its peak: the moronic and meaningless case of only one response category!

the adaption of an established proposal is to try to systematize and streamline our present scheme. Besides the superficially rather good evaluation results, the advantage with this second option is that our dialogue coding draft is already designed to be up to the mark of the object identification games played regarding certain aspects. Note, however, that we will have some suggestions to make that amount to a refinement of the present scheme on systematic grounds in section 4 below. If respective improvements were implemented, it would be interesting to determine the reliability results for the refined scheme and compare them with those for the current one. Indeed, it must be taken into account that the results for the refined scheme might be worse than those obtained at the moment. However, this would not necessarily speak against the changes if, for example, the usage of the categories would be more balanced. At the moment, we are not certain of how we will choose; maybe the proof of the pudding will be in the eating!

The results of our investigations pertaining inter-rater agreement can be summarized as follows: some of the more theoretically-loaded category ratings show a distinctively lesser degree of agreement as compared to the rather "theoretically innocent" ones. In opposition, the categorical dialogue move scheme shows impressively good results, but it would be interesting to see whether it can up live up to this quality in a more complex setting and, possibly, after the introduction of systematic refinements with regard to the move types that have been used overly frequently. All in all, this may seem to be a somewhat pessimistic perspective in the face of comparably good results—however, we think that it is a productive one indeed, since it leaves us with an agenda on where to improve on the framework in the nearer future.

4 Towards Theory-oriented Annotation

The annotation scheme that underlies our raters' decisions has been described to some degree of detail in section 2.2 above. Afterwards, we have come to present the reliability results for the scheme and have commented on these results, both in the preceding chapter. Now, despite the rather good results concerning its dialogue parts, the scheme is a tool that has been developed to cope with a certain task: in our case its basic purpose was mainly explorative in nature. In addition, during the course of its application, some shortcomings got apparent that had to be remedied in an, admittedly, *ad hoc* fashion. Therefore, the DEIKON coding is an instantiation of what we would like to call a *problem-oriented annotation* here and which we understand as a way of coding that is done with a specific and restricted aim in mind. Further, with problem-oriented annotations the coding scheme itself can well be object to modification. This we regard as opposed to *theory-oriented annotation*, where the categorical inventory and the coding instructions of the coding scheme are derived from an underlying theory and can not be altered in annotation sessions.[1] The distinction between problem- and theory-oriented annotation is not an established but an original one that we introduce here. However, it bears some resemblance to a distinction that has been introduced in a precursor to this report [Kühnlein and Stegmann, 2003] which runs between *bottom-up* and *top-down annotation*. In this respect, it seems that problem-oriented annotation shows some affinity to the bottom-up strategy and, in turn, theory-oriented annotation resembles the description of the top-down perspective. However, the *foci* seem to be slightly different. This notwithstanding, it might be interesting to speculate whether a step towards a theory-oriented perspective in the future might lend itself to a change of the annotation software used, cf. the respective arguments in [Kühnlein and Stegmann, 2003].

Indeed, we think that a serious annotation project has to build on a good underlying theory. In the following, we want to point to some defects that we have detected in our scheme and relate them to the potential bearing they might have concerning reliability measures.

To begin with, depending on whether the scheme used tends more towards a problem-oriented or a theory-oriented perspective, this might influence the outcome of subsequent agreement measures. Just one example: theory-oriented coding schemes seem to lend themselves to the use of more annotation tiers, since they are designed to cover all relevant aspects of a certain phenomenon—an aim which a problem-oriented scheme does not need to account for due to its focus on certain, restricted aspects. But then, increasing the number of response categories, *ceteris paribus*, also increases the possibilities

[1] Of course, in practice nearly any problem-oriented annotation will depend on some theory. For example, part of speech tagging relies on grammatical categories, which are technical terms. Therefore, the distinction between problem- and theory-oriented approaches to annotation is an idealized one.

for disagreement. This notwithstanding, with a theory-oriented scheme, more disagreement need not necessarily be taken to indicate a deterioration of the goodness of the scheme, if the less good results are achieved against the background of a generally more adequate use of available categories, i. e., a more elaborate scheme.[2] However, the alteration of agreement values in comparing problem-oriented and theory-oriented schemes can go in both directions. For example, a more elaborate and systematic repertoire of response categories might also pave the way for the formulation of clearer annotation instructions, which in turn might lead to better reliability results. There are a couple of remarks that have to be made concerning our problem-oriented coding scheme seen from a more theory-oriented perspective. They will be mostly related to the dialogue coding parts of the annotation scheme. However, it should be noted that certain points of critique concerning other parts of the scheme have already been expressed in the appropriate parts of section 2.2 above.

One of the challenges the DEIKON scheme has to master is the integration of gestures. Now our point to be made here concerns the fragmentary nature underlying the multimodal extension of the employed move types. Since both subjects, Instructor as well as Constructor, could make use of two communication channels, viz. gestural and vocal, as well as their combination, there are three possibilities how a dialog contribution can be realized: purely verbal, purely gestural, or multimodal. A coding scheme that aims towards the role of pointing in task-oriented dialogue should reflect this triple systematically. However, the DEIKON scheme consistently neglects one of them! Although there is a move type label called `check-back` covering multimodal check-backs and a move type label `act/check-back` that is used for feedback performed with a gesture exclusively, a category like `verb/check-back` (we don't want to argue about denomination) is missing which might be reserved for purely verbal check-backs. The same holds, e. g. for `demonstration` and `complex demonstration`. The reason for this lack of categories is that the data provide no cases of purely verbal check-backs (similarly not for purely gestural demonstration). Though this might be reasonable for problem-oriented approaches, it is a shortcoming in theory-oriented ones. Furthermore, the bare appending of a category (whether it is used or not) influences the calculation of inter-rater agreement statistics. The more categories are involved, the less is the probability for agreement by chance. This is easily seen recognizing that AC_1's chance term is limited by $\frac{1}{k}$ for k categories. Thus, restricting the inventory of classification labels to those that actually get used increases the result of the statistics, since it decreases the estimation of the chance term!

Besides nagging about more or less quantitative deficiencies, we also have to complain

[2] Imagine a scheme with, say, only two categories and, further, that one of the two categories is used very often. Now, a finer-grained scheme might split that category in several finer-grained (sub-)categories. For example, let's say that the "repair" category is the one that has been used very often. Then we might choose to have "repair-of-clarification", "repair-of-assertion", "repair-of-directive" and further comparable sub-repair-types instead of the one general category. Such refinements might lend themselves to more disagreements at face value, since raters who agree on the general category may still disagree with respect to the correct specialization. However, such disagreement may be a price worth paying in exchange for a finer-grained, more elaborate and theoretically adequate annotation scheme.

about qualitative ones. Recall from section 2.2 that we distinguish between `repair` and `repair-of-clarification`. Apparently, both move type labels are settled on different levels: whereas the latter is a special kind of repair, the former is a generic one. Such unbalanced partitioning of move type labels not only complicates the formulation of coding instructions but also raises difficulties in applying them.

One last suggestion concerns the dialogue game level. Currently, the scheme focuses solely on object identification games. For these games, the underlying intentions are related to the main task that we have set for our subjects. However, it has been observed that lots of meta-communicative issues are raised and settled during the realization of the identification tasks, compare also the explicit usage of corresponding entities in the current dialogue move scheme, e. g., `repair`, `clarification`, and `repair-of-clarification`. Therefore, it would be most appropriate to include corresponding meta-communicative sub-dialogue games such as “meaning identification” and/or “repair” games on an appropriate annotation tier. Of course, this would complicate matters considerably, since such games will be embedded within object identification games or may run concurrently with them. Nevertheless, such annotations would be necessary in order to investigate the intricate details of meta-communicative exchanges in a multi-modal domain.

For the sake of completeness we have to confess that some naming conventions seen in retrospective are infelicitous at best. Originally, they were designed to provide a systematic basis for XSLT-related processing means: therefore, the speech-related tiers are preceded by a `speech.`-prefix while the gesture-related tiers have a `gesture.`-prefix. In the light of this strategy, it would have been desirable to provide a uniform `dialogue.`-prefix for the `move.type` and `game.type` tiers, since both belong to the same domain of investigation. Furthermore, the present `.type`-suffix is redundant for both tiers, since it adds no information that could be made use of for means of discrimination or the likes. Besides an improvement on the system, respective changes might come in handy for dialogue-specific processing and analysis means in the nearer future.

There is a further naming convention point to be made, this time with regard to the repertoire of dialogue moves: it pertains to the chosen names for the `demonstration` and `complex-demonstration` elements. From the point of view of the linguistic elements contained, the related moves do not comprise demonstrations in a narrow sense, e. g. they do not contain demonstrative pronouns as might be expected due to their names. Indeed, originally we had expected to find utterances as “that yellow bolt” or “this ↘ yellow bolt” to be performed by our subjects. However, only definite descriptions have been found, e. g., utterances as “the yellow bolt” or its multi-modal counterpart “the ↘ yellow bolt”. Hence, in the light of the empirical data collected it seems to be more appropriate to exploit labels like `definite-description` and `complex-definite-description`. We could speak of a “demonstration” in a wider sense if such deictic elements as bodily orientation, accompanying eye gaze and/or other means of exploitation of the situative context had our main attention with regard to the phenomena under discussion. However, this does not seem to be the case, since those elements are not part of the instructions underlying the annotation yet. Therefore, there seems to be no reason to indicate such matters explicitly in the naming of the respective elements. It might be objected that if a gesture accompanies the uttering, we have a case of demonstration. However, the occurrence of a gesture is already indicated by means of the `complex`-prefix.

Bibliography

Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, pages 249–254, 1996.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–32, 1997.

Herbert H. Clark. *Using Language.* Cambridge UP, 1996.

Ann Copestake. *Implementing Typed Feature Structure Grammars.* CSLI Lecture Notes (No.110). CSLI Publications, 2002.

Mark G. Core and James F. Allen. Coding dialogues with the DAMSL annotation scheme. In David Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park California, 1997. American Association for Artificial Intelligence.

Corpus. *Wir bauen also jetzt ein Flugzeug*. Konstruieren im Dialog, Arbeitsmaterialien. Interaktion sprachlicher und visueller Informationsverarbeitung. Collaborative Research Center SFB 360 “Situated Artificial Communicators”, Bielefeld University.

Barbara DiEugenio and Michael Glass. The kappa statistic: a second look. *Computational Linguistics*, 30(1), 2004.

Kilem Gwet. *Handbook of Inter-rater Reliability.* STATAXIS Publishing Company, 2001. URL `http://www.stataxis.com/`.

Charles Hartshorne and Paul Weiss, editors. *Collected Papers of Charles Sanders Peirce*, volume II. Belknap Press, 1967. repr. from 1932.

Bernhard Jung. *Wissensverarbeitung für Montageaufgaben in virtuellen und realen Umgebungen.* PhD thesis, Bielefeld University, Faculty of Technology, Artificial Intelligence Group, 1996.

Stefan Kopp and Ipke Wachsmuth. Synthesizing multi-modal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.

Alfred Kranstedt, Stefan Kopp, and Ipke Wachsmuth. MURML: A multimodal utterance representation markup language for conversational agents. In *Proceedings Workshop Embodied Conversational Agents*, Bologna, Italy, 2002. First International Joint Conference on Autonomous Agents & Multi-Agent Systems.

Peter Kühnlein and Jens Stegmann. Empirical issues in deictic gestures: Referring to objects in simple identification tasks. Technical Report 2003/03, CRC 360 “Situated Artificial Communicators”, Bielefeld University, 2003.

Peter Kühnlein, Manja Nimke, and Jens Stegmann. Towards an HPSG-based formalism for the integration of speech and co-verbal pointing. In Jürgen Streeck, editor, *Gesture: The Living Medium. Proceedings of the first congress of the International Society for Gesture Studies (ISGS).* University of Texas at Austin, June 5-8, 2002 2003. URL `http://www.utexas.edu/coc/cms/International_House_of_Gestures/Conferences/Proceedings/Contents/List_of_Papers.html`.

Peter Kühnlein, Alfred Kranstedt, and Ipke Wachsmuth. Deixis in multi-modal human computer interaction: An interdisciplinary approach. In A. Camurri and G. Volpe, editors, *Gesture-based communication in human-computer interaction*, number 2915 in Lecture Notes in Artificial Intelligence, pages 112–123, International Gesture Workshop 2003, Genua, Italy, 2004. Springer: Berlin, Heidelberg. revised papers,.

Andy Lücking and Hannes Rieser. Empirische Untersuchung von Zeigegesten in experimentellen Settings. Manuscript, June 2004.

Andy Lücking, Hannes Rieser, and Jens Stegmann. Statistical support for the study of structures in multi-modal dialogue: *Inter*-rater agreement and synchronization. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog '04—Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue*, pages 56–63, Barcelona, 2004. Department of Translation and Philology, Universitat Pompeu Fabra.

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

Jan-Torsten Milde and Ulrike Gut. The TASX-environment: an XML-based corpus database for time aligned language data. In *Proceedings of the IRCS Workshop on linguistic databases, Philadelphia*, 2001.

Carl J. Pollard and Ivan A. Sag. *Information-based Syntax and Semantics, Vol. 1.* Number 13 in CSLI Lecture Notes. CSLI Publications, Stanford University, 1987. Distributed by University of Chicago Press.

Carl J. Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago, 1994.

Willard Van Orman Quine. *Word and Object.* Studies in Communication. MIT Press, 1960.

Hannes Rieser. A unified account for gesture meaning and expression meaning in simple reference games. Manuscript, 2001.

Hannes Rieser. Pointing in dialogue. In Jonathan Ginzburg and Enric Vallduví, editors, *Catalog '04—Proceedings of the Eighth Workshop on the Semantics and Pragmatics* *of Dialogue*, pages 93–100, Barcelona, 2004. Department of Translation and Philology, Universitat Pompeu Fabra.

Ivan Sag and Thomas Wasow. *Syntactic Theory—A Formal Introduction*. CSLI, 1999.

Jens Stegmann. Thesen zur Theorie und Empirie einer multi-modalen Grammatik. Manuscript, 2004.

Jens Stegmann and Andy Lücking. Assessing reliability on annotations (1): Theoretical considerations. Technical report, Universität Bielefeld, SFB 360, Projekt B3, 2005.

Ludwig Wittgenstein. *The Brown Book*. Blackwell, 1958.