Statistical Support for the Study of Structures in Multi-Modal Dialogue: *Inter*-Rater Agreement and Synchronization

Andy Lücking **Hannes Rieser** **Jens Stegmann**
SFB 360 “Situated Artificial Communicators”, B3
Bielefeld University
{andy.luecking|hannes.rieser|jens.stegmann}@uni-bielefeld.de

Abstract

We present a statistical approach to assess relations that hold among speech and pointing gestures in and between turns in task-oriented dialogue. The units quantified over are the time-stamps of the XML-based annotation of the digital video data. It was found that, on average, gesture strokes do not exceed, but are freely distributed over the time span of their linguistic affiliates. Further, the onset of the affiliate was observed to occur earlier than gesture initiation. Moreover, we found that gestures do obey certain appropriateness conditions and contribute semantic content (“gestures save words”) as well. Gestures also seem to play a functional role wrt dialogue structure: There is evidence that gestures can contribute to the bundle of features making up a turn-taking signal. Some statistical results support a partitioning of the domain, which is also reflected in certain rating difficulties. However, our evaluation of the applied annotation scheme generally resulted in very good agreement.

1 Introduction

In ordinary face-to-face communication, people make use of both speech and non-verbal gesticulation. No reductive relationship holds between these modes of communication in either direction. This assumption is in accordance with empirical work, e. g. in psycholinguistics (McNeill, 1992, e. g.), as well as with philosophical considerations, mainly about reference and demonstration (Wittgenstein, 1958; Peirce, 1965). Hence, we take it as a truism that accounts of dialogue must be extended to include a treatment of gesture.

Empirical investigations of multi-modal dialogue comprising gesture and speech can pursue at least two interests: First, one wants to know how speech and pointing gestures are related to each other, especially whether the information from the auditory and from the visual channel synchronizes. Here the focus is on relations within individual dialogue moves. We call this ‘*intra*-move synchronization’. Secondly, a similar interest exists concerning pointing gestures and exchanges of turns, where the question is how speech and gesture of one speaker are related to the gestures and the speech of his addressee and *vice versa* (‘*inter*-move synchronization’). Here the focus is on relations between different dialogue moves within one dialogue game.

The distinction between *intra*- and *inter*-move synchronization reflects different research lines that have been pursued in recent years. Psycholinguistics serves as an illustrative example here. One point of reference is the body of work in gesture studies that builds on McNeill (1992), whose main empirical focus is on the relationships holding among gestures and speech within utterance units. On the other hand, much current work in dialogue theory centers on issues that are intimately connected with coordination among language users, e. g. building upon the *joint actions* framework of Clark (1996); but see also the notion of alignment in (Pickering and Garrod, in press).

Our investigation is based on original empirical studies. The task we set for our subjects involved the choice of referents from a restricted domain, see figure 1 and figure 2. They had to negotiate or to align reference using dialogue games of a certain type. In order to get results showing relations obtaining between gesture and speech in dialogue, we applied descriptive and analytical statistical methods to the time-based annotation stamps of suitable dialogue data. Such statistical analysis is pointless, of course, unless the employed annotation scheme isn't evaluated and confirmed to be reliable.

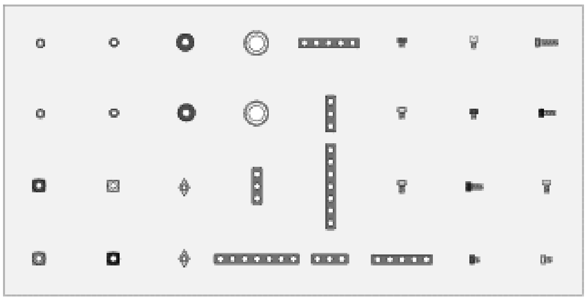


Figure 1: The pointing domain (form cluster), taken from (Kühnlein and Stegmann, 2003).

Accordingly, we present our study as follows: First, we set the stage with a description of the annotation of the empirical data (section 2). We then report on assessing of *inter*-rater agreement on our annotation scheme (section 3). In section 4 we present the results of further empirical investigation, mainly concerned with synchrony. We conclude the paper with a summary of our findings and a discussion of those topics that might be explored in further studies (section 5).

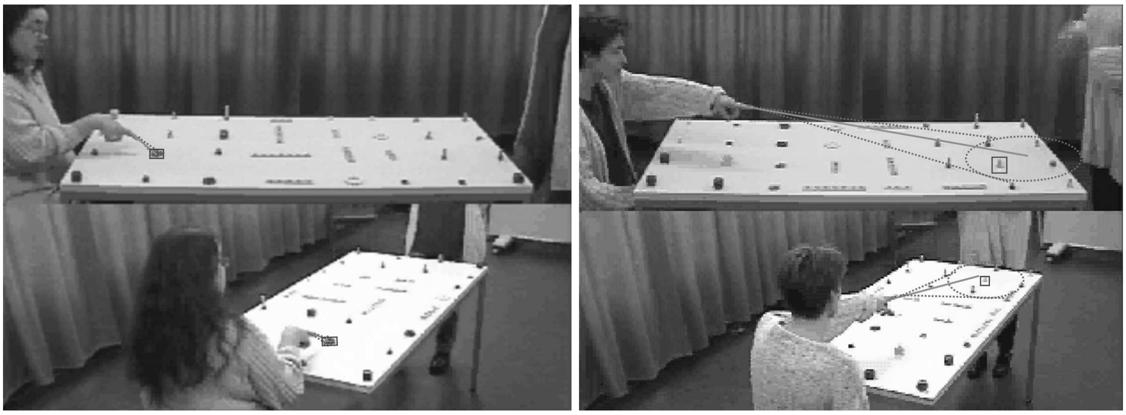
One last word of caveat: note, that our empirical studies are preleminary in the sense that only some variables have been controlled. This is due to the fact that the studies had not been conducted with issues of precise statistical hypothesis testing in mind. However, the results reported here are reasonably robust and will be reproducible in more carefully controlled experiments (see section 5).

2 Annotation of simple reference games

The analysis of our corpus of digital video data is based on an annotation with the TASX-ANNOTATOR software package[1] (Milde and Gut, 2001) which allows for the pursuit of an XML-based bottom up approach. Since the annotation data are stored in XML format, the extraction of the relevant information for purposes of statistical analysis could be realized *via* XSLT script processing straightforwardly. Details connected with the empirical setting and principles of annotation are laid out in (Kühnlein and Stegmann, 2003).

Figure 3 is a screenshot from a TASX annotation session that exemplifies the annotation scheme applied in score format. The set of annotation tiers includes a transcription of the agent's speech at word level (*speech.transcription*) and a classification of the dialogue move pursued (*move.type*). The annotation of deictic gestures follows the framework established by McNeill (1992). A gesture token has three phases: wrt pointing gestures the maximally extended and meaningful part of the gesture is called *stroke*, and *grasping* if an agent grasps an object. Stroke or grasping are preceded by the *preparation* phase, that is the movement of the arm and (typically) the index finger out of the rest position into the stroke position. Finally, in the *retraction* phase the pointer's arm is moved back into the rest position. We presume that pointing gestures serve one of two semantic functions: they uniquely pick out an object (*object pointing*) or merely narrow down the region in which the intended object lies (*region pointing*). In order to clarify this distinction, in figure 2 an occurrence of each gesture function is shown. The extension of pointing gestures is modelled with a pointing cone. Subfigure 2(b) depicts a case of region pointing, where several objects are located in the conic section of the pointing cone and the table top. In addition, the extension of the index finger does not meet the object in question. Against this, in object pointing the object is unequivocally singled out, i. e. it is the only object within the conic section, see subfigure 2(a). Seeing the "fuzziness" of pointing gestures as antic-

[1] It can be obtained at `http://tasxforce.lili.uni-bielefeld.de/`.



(a) Object pointing

(b) Region pointing

Figure 2: Pointing cones. The extension of the index finger is indicated with a line, the pointing cone is indicated by dotted lines, and the box frames the intended object.

ipated by Quine's (1960) thesis of the indeterminacy of reference, the philosophical stance taken here can be labelled as *neo*-Peirce-Wittgenstein-Quinean (Rieser, 2004). The distinction between object and region pointing is captured on the *gesture.function* tier.

All tiers are specified for instructor and constructor, i. e. the respective tier names have an *inst.* or *const.* prefix, cf. figure 3.

To get a better grip on the kind of data we are concerned with, the speech portions of the sample dialogue from figure 3 were extracted and are reproduced below.

(1) Inst: The wooden bar
[pointing to object1]

(2a) Const: Which one?

(2b) This one?
[pointing to object2]

(3a) Inst: No.

(3b) This one.
[pointing to object1]

(4) Const: This one?
[pointing to object1 and grasping it]

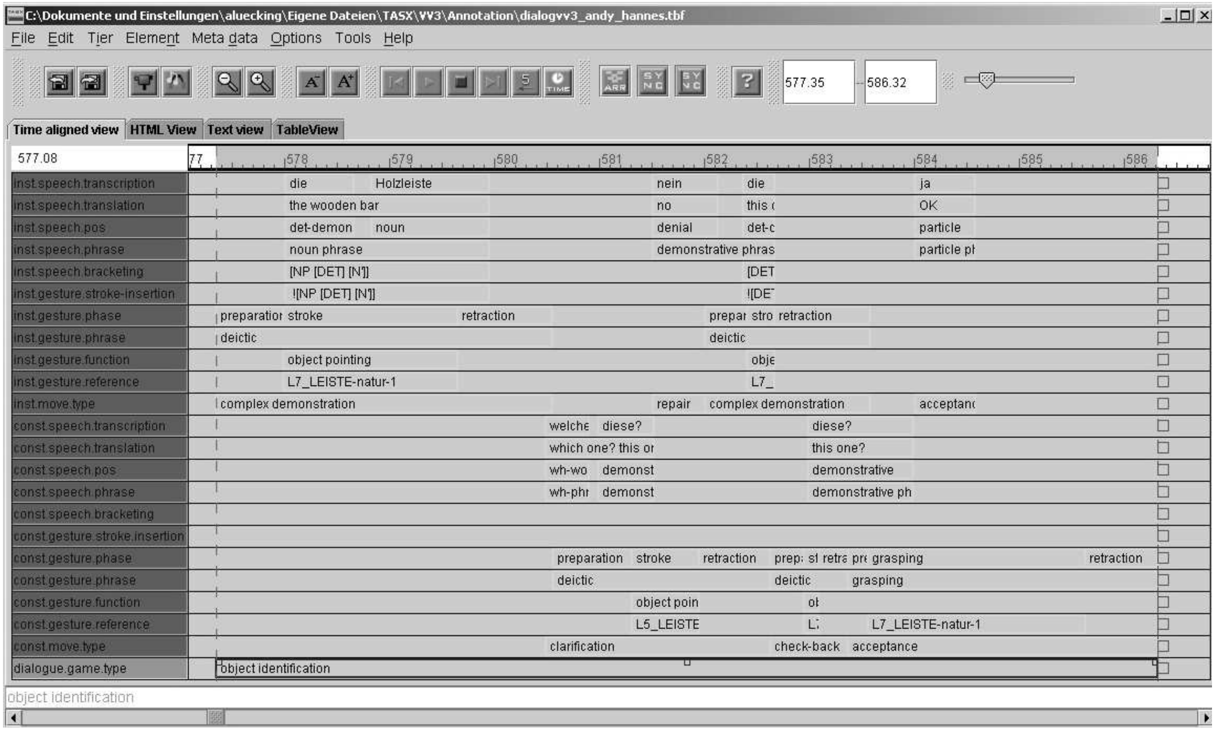
(5) Inst: O.K.

We have the dialogue move of a *complex demonstration* of Inst's in (1) here, followed by a *clarification* move involving a pointing of Const's (2a, 2b). Inst produces a *repair* (3a), followed by a new *complex demonstration* move (3b) to the object she had introduced. Then we have a new *check-back* from Const (4) coming with a pointing and a grasping gesture as well as an acceptance move by Inst (5). The whole game is classified as an *object identification game*. The following events from different agents' turns overlap: (2b) and ((3a) and (3b)); (3b) and (4).

3 Reliability of the Annotation Scheme

Annotation-based projects must decide on the appropriateness of the annotation scheme. The standard way to handle this is using a bag of statistical methods that goes under the heading of *inter-rater agreement* or *inter-rater reliability*. Basically, the underlying idea is that of conducting a test on the results of raters who have annotated the same set of data. Different aspects of reliability (stability, reproducibility, and accuracy) go with different test designs (test *vs* retest, test *vs* test, and test *vs* "gold standard") and different *foci* of measured error (*intra*-observer, *inter*-observer, and deviation from norm) (Krippendorff, 1980). We are concerned with the second aspect of reliability (reproducibility, test *vs* test, *inter*-observer) here, since we have evaluated our annotation scheme comparing two raters' codings of the same video data.

In dialogue research the most widely known proposal concerning measures of *inter*-rater agreement is (Carletta, 1996) who argues in favor of the *kappa* statistics. However, there are serious problems associated with its interpretation, cf. (Feinstein and Cicchetti, 1990) on kappa paradoxes.


C:\Dokumente und Einstellungen\aluecking\Eigene Dateien\TASX\VV3\Annotation\dialogvv3_andy_hannes.tbf
File Edit Tier Element Meta data Options Tools Help
577.35
586.32
Time aligned view
HTML View
Text view
TableView
577.08
77
578
579
580
581
582
583
584
585
586

Tier	Elements
inst.speech.transcription	die Holzleiste nein die ja
inst.speech.translation	the wooden bar no this (OK
inst.speech.pos	det-demon noun denial det-c particle
inst.speech.phrase	noun phrase demonstrative phras particle pł
inst.speech.bracketing	[NP [DET] [N']] [DET
inst.gesture.stroke-insertion	![NP [DET] [N']] ![DE
inst.gesture.phase	preparatior stroke retraction prepar stro retraction
inst.gesture.phrase	deictic deictic
inst.gesture.function	object pointing obje
inst.gesture.reference	L7_LEISTE-natur-1 L7_
inst.move.type	complex demonstration repair complex demonstration acceptanc
const.speech.transcription	welche diese? diese?
const.speech.translation	which one? this or this one?
const.speech.pos	wh-wo demonst demonstrative
const.speech.phrase	wh-phr demonst demonstrative ph
const.speech.bracketing	
const.gesture.stroke.insertion	
const.gesture.phase	preparation stroke retraction prep: st retra pre grasping retraction
const.gesture.phrase	deictic deictic grasping
const.gesture.function	object poin ol
const.gesture.reference	L5_LEISTE L; L7_LEISTE-natur-1
const.move.type	clarification check-back acceptance
dialogue.game.type	object identification

object identification

Figure 3: Annotation of a more complex Dialogue Game.

The point is that in the calculation of kappa the term representing the proportion of agreement by chance is systematically overestimated. Therefore, where appropriate with respect to the type of data involved, we pursue an alternative proposal based on the methodological framework of Gwet (2001), i. e. his *AC1* statistics. The latter—more adequately chance-corrected—coefficient is appropriate with respect to data resulting from a *type-ii* measurement on nominal-scale niveau.[2] Concerning judgments on magnitude scale niveau, which are usually classifiable as being of *type-i*, we use well-established conventional techniques, mainly correlation analysis. All calculations were implemented making use of the statistical programming environment R (R Development Core Team, 2003)[3].

Our *type-i* annotation data on a magnitude scale are the time-stamps for words and gestures, i. e. the points in time when words begin or end, and the start or end times of the gesture phases. In the TASX-ANNOTATOR a time bar is incorporated and synchronized with the video, so that a mark on the *speech.transcription* tier, say, at 201.4 seconds, means that the word in question starts at second 201.4 of the respective entire videotaped session. Since performing a gesture is a continuous action, the coding of gesture phases splits it into three parts where the end time of the preceding phase is identical with the start time of the following one. For example, the end of the preparation simultaneously marks the start of the stroke. The correlation of those time-based annotations was calculated over 108 words and 25 gesture occurrences using the Pearson product-moment correlation coefficient r. The outcomes are given in table 1. Despite almost perfect values of nearly 1, there is need for a closer look, since this result is influenced by the strict linearity of the underlying time scale. We transformed linear measurement data into nominally scaled data because of the category of stroke insertion, which is derived from allocating the stroke element's time in-

[2] *Type-ii* measurements are those, where the process leading to the measured datum is not well understood. Comparably well-understood measurements go by the name of *type i*. We will overload the term to refer to respective data, where appropriate.

[3] http://www.r-project.org.

	preparation	stroke		retraction	word boundaries	
	start	start	end	end	start	end
r	0.9999999	0.9999999	0.9999998	0.9999976	0.9999999	0.9999999

Table 1: Results for the correlation of gesture and word boundaries.

terval relative to the part of speech portions. This means, basically, a projection from temporally extended entities onto a sequence of symbols on, say, a modality-neutral representation at roughly word level, which could be fed into a parser. Essentially, we abstract away from exact timing—only the relative order remains, cf. example sentences 1 and 2 below, where ↘ symbolizes gestural stroke.

(1) ↘ the wooden bar (2) the ↘ wooden bar

Resulting in nominally scaled data, the agreement regarding stroke insertion could be calculated using AC1, leaving us with a value of "merely" 0.73, which still can be regarded as good agreement. However, this result reveals that minor deviations in determining the boundaries of parts of speech and gesture phases can make a difference for the exact placement of the stroke.

One main concern was to assess whether the distinction between object pointing and region pointing is a concept reproducible by different raters. Being a nominal response category resulting from a *type-ii* measurement, the degree of correlation in classifying gesture functions was calculated using AC1. With a value of 0.4842 that is based on the judgment of 56 gesture occurrences, the agreement has to be classified as being fair at best. This shows that there are different habits in judging gestures as being related to object or region, which, in turn, indicates that either a clear-cut empirical category is lacking, or that the two-dimensional video data are not good enough to admit of this categorization.

Nevertheless, there was close agreement among raters concerning certain regions of the pointing domain. The domain of the reference games can be partitioned into three regions, according to the distance measured from the instructor, cf. figure 1. The two leftmost columns form the proximal region, the two rightmost columns the distal region, and the remaining four columns are called the mid-range region. Observe now that there is nearly perfect correlation with respect to the categorization of pointing to objects located in the proximal or distal regions. Dissent arises wrt pointing into the mid-range area. This shows that reliability of assignment of gesture functions is conditioned by the relative position of the objects that are referred to by the instructor.

Being interested in the dialogue structure of reference games, we also checked the reliability of our dialogue move annotation scheme. This was carried out computing the AC1 separately for instructor and constructor moves. The highly schematic instructor moves form a recurrent pattern that could be judged fairly consistently in the annotations of both observers ($N = 92, \mathrm{AC1} = 0.9$). Agreement diminished when concerned with the more variable constructor moves ($N = 65, \mathrm{AC1} = 0.795$).

4 Empirical findings

Gestures contribute to the content of communicative acts rather than being mere emphasis markers. This hypothesis can be substantiated by findings related to the semantic, the pragmatic, and the discourse level. On the semantic level, gestures contribute content that otherwise would have to be cast into clumsy verbal descriptions, thus making communicative acts more efficient. We found strong evidence for this in comparing the number of words used in referential NPs without pointing gesture occurrences (hereafter DDs, for *definite descriptions*) with NPs that come with pointing gestures (CDs, short for *complex descriptions*). A *t*-test was carried out on the number of words used in 65 CDs *vs* that in 74 DDs, resulting in a (highly) significant difference ($t = 6.22, p = 5.696 \cdot 10^{-9}, \alpha = 0.05$), cf. figure 4. This result can be couched into the slogan "Gestures save words!".

A related hypothesis was that the time the con-

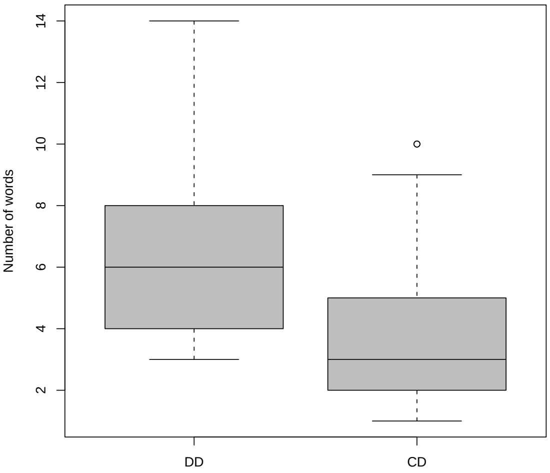


Figure 4: Boxplot displaying the number of words in CDs and in DDs.

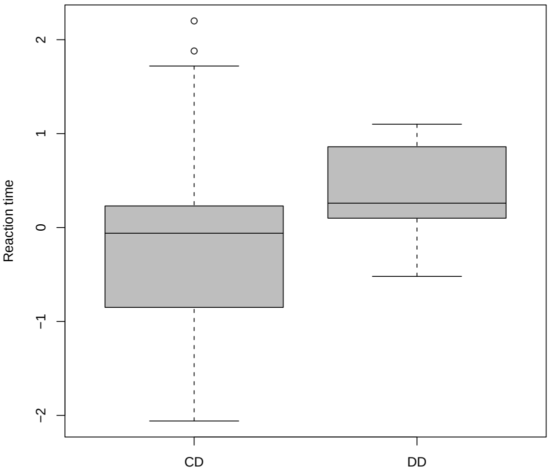


Figure 5: Boxplot for Const's reaction times (in seconds) following Inst's CDs and DDs.

structor needs to interpret the instructor's reference (reaction time) will be less after a CD than after a DD. The pointing gesture can be seen as guiding the constructor's eye towards the intended object—or at least towards a narrow region where the object is located—and thus as shortening the constructor's search effort. To assess this point, we calculated 48 (39 CDs and 9 DDs) differences between the start time of the constructor's move and the end time of the instructor's preceding referring act. A subsequent *t*-test did not result in a significant difference ($t = -1.4, p = 0.166, \alpha = 0.05$), but there seems to be a tendency for shorter reaction times after CDs, cf. figure 5.

What might have prevented a significant outcome was the fact that some objects are unique and therefore more salient, e. g. there is only one yellow cube (as opposed to several yellow bolts), so that the constructor could quickly spot such objects when directed with appropriate DDs only. In addition, the constructor may have used the instructor's gaze as a guiding device.

Moving from semantic to pragmatic issues, we also tried to find out whether there are contextual conditions constraining the use of gestures. This was defined in terms of frequencies of DDs *vs* CDs utilized to refer to objects in different columns of the pointing domain—that is, basically, wrt their distance as seen from the instructor. What is at stake here is whether the asymmetry that seems to be revealed in the bare data—comparethe plot depiction in figure 6—could be statistically validated; with DD's frequency peaks in the *periphery* (that are columns 1 plus 2 and 7 plus 8, or in terms introduced earlier, the union of the proximal and the distal region) and CD's frequency peaks in the *center* (the mid-range region, columns 3 to 6), there should be a bias to demonstrate objects in the middle of the domain using pointing gestures, whereas objects located in peripheral areas should be referred to only verbally.

There are two questions that have to be distinguished: First, is there a difference in the proportions of CDs *vs* DDs wrt the peripheral, resp. the center, region? Secondly, is there a difference in the proportions of CDs, resp. DDs, wrt the regions? To assess the second point the frequencies of peripheral and center CDs were compared using a χ^2-test, resulting in a significant outcome ($N_{\text{peripheral}} = 24, N_{\text{center}} = 41, \chi^2 = 7.8769, p = 0.005, \alpha = 0.05$). The comparison of the frequencies of DDs modelled through periphery and center yields an analogous result ($N_{\text{peripheral}} = 46, N_{\text{center}} = 28, \chi^2 = 8.7568, p = 0.003, \alpha = 0.05$). As regards the first issue, comparing the proportions of CDs *vs* that of DDs to refer into the peripheral (and likewise the center area), we get significant outcomes, too (for periphery: $N_{\text{CD}} = 24, N_{\text{DD}} = 46, \chi^2 = 13.8286, p = 0.0002, \alpha = 0.05$; for center: $N_{\text{CD}} = 41, N_{\text{DD}} = 28, \chi^2 = 4.8986, p = 0.027, \alpha = 0.05$). Thus, the relative distance of the object in question to the instructor is a contextual factor for the choice of the mode of reference to that object. It is noteworthy that the partition of the reference domain imposed by the ratings of gesture function coincides with that of capturing the CD/DD-asymmetry.

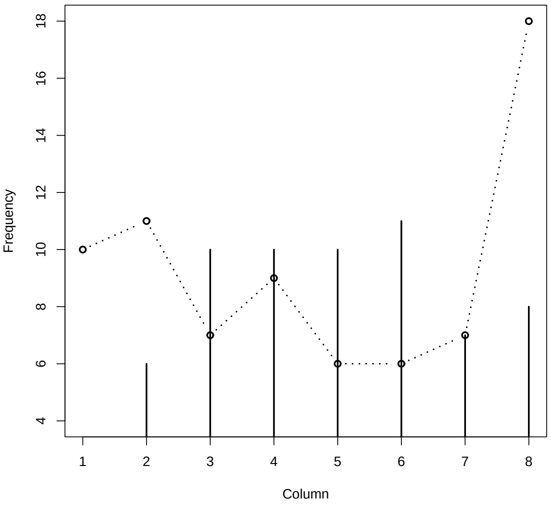


Figure 6: Plot for the modes of reference modelled by the eight columns of the reference domain; the bars depict the frequency distribution of CDs over the columns, the dashed line that of DDs.

At the beginning of this paper, a distinction was made between *intra*- and *inter*-move synchronization at the dialogue level. As regards *intra*-move synchronization we accounted for the temporal relations holding between gesture phases and escorting utterances. Above all, we focused on two synchronization effects, namely *anticipation* and *semantic synchrony* (McNeill, 1992, pp. 25-26, p. 131). The semantic synchrony rule states that gesture and speech present one and the same meaning at the same time (McNeill's "idea unit"). Anticipation refers to the temporal location of the preparation phase in relation to the onset of the stroke's co-expressive portion of the utterance. This rule states that the preparation phase precedes the linguistic affiliate of the stroke. Table 2 summarizes the descriptive statistics ($N = 25$).[4] Note, that we take the verbal affiliate to be the complete denoting linguistic expression, i. e. a possibly complex noun phrase. Row P gives the values for the start of the preparation phase relative to the onset of the first word of the noun phrase.Contrary to McNeill (1992, p. 25, 131), we found that the utterance usually starts a little before the initiation of the gesture (compare the positive mean value in table 2). This seems to contradict anticipation, given the way we operationalised McNeill's concept of the verbal affiliate or the idea unit. Similarly (compare the mean value in row R), the stroke ends (or the retraction starts) normally around 0.5 seconds before the end of the affiliate. Together with an average beginning of the stroke around 1 second after the onset of the utterance (mean for row S) this shows, that the prototypical stroke does not cross utterance boundaries. This is as to be expected in the light of McNeill's semantic synchrony rule. Note, however, that some extreme tokens (compare respective min. and max. values in table 2) were observed that seem to contradict the McNeill regularities, cf. (Kühnlein and Stegmann, 2003).

	Min.	Mean	Max.	Std. Deviation
P	-0.8	0.3104	4.68	1.0692
R	-0.86	0.564	3.38	0.89
S	-0.02	1.033	5.54	1.128

Table 2: *Intra*-move synchronization of preparation (P), retraction (R), and stroke (S).

Concerning *inter*-move synchronization one point of interest was the alignment of the end of Inst's preparation phase with Const's retraction phase. The resulting values, given in table 3, show that there is gap of around 0.3 seconds at aver-

Min.	Mean	Max.	Std. Deviation
-2.06	0.29	3.46	1.27

Table 3: *Inter*-move synchronization of Const's retraction and Inst's preparation.

[4] The different rows were calculated as follows: (P) preparation$_{\text{start}}-$ speech$_{\text{start}}$, (R) speech$_{\text{end}}-$ retraction$_{\text{start}}$, and (S) stroke$_{\text{start}}-$ speech$_{\text{start}}$.

age. But the comparatively large values for the range (the span between the maximum and minimum values observed) and the standard deviation suggest that simply averaging the results camouflages a great deal of dispersion. A look at the dialogue video data reveals roughly two different sources for the resulting large and small values. If the object referred to lies within Const's reach, his initiation overlaps with Inst's retraction, indicating that the retraction phase contributes to a turn-taking signal. If the object referred to lies at the opposite side of the table Const first has to move around the table which delays initiation of her gesture.

5 Prospectus

As pointed out in the course of this paper, there are some rough edges in the employed annotation scheme as well as findings that can't be accounted for properly as yet. Accordingly, the top of our agenda includes experiments suitably designed to determine (or at least approximate sufficiently) the topology of the pointing cone. Such findings, we hope, will improve the classification of gesture functions and shed some light on the role the partitioning of the domain plays in the manner of how reference is established.To streamline our coding of move types we will hook up to some already established annotation scheme. At the time being, the one that seems to be most appropriate for our kind of data is the HCRC coding scheme (Carletta et al., 1996), which has to be augmented to capture pointing gestures. A third topic that could be fruitfully invistigated concerns the interaction of speech, gesture and gaze, which opens the door to truly *multi*-modal dialogue. As remarked above, the constructors in our settings might have used instructors' eye movement as an information source to find out the location of the object in question. As regards *intra*-move synchronization, we found a variety of temporal relationships that exceeds by far what was to be expected in the light of the current literature. In addition, we found surprising variability with respect to *inter*-move synchronization. Especially the frameworks aiming at a phenomenological account of gestures (mainly based on *iconics*) do not capture the structural flexibility of deictic gestures. A more promising direction to approach pointing and grasping in dialogues should perhaps be based on rigid semantics and underspecification approaches, cf. (Rieser, 2004).

References

J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1996. HCRC dialogue structure coding manual. Technical Report TR-82, University of Edinburgh.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 2(22):249–254.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

A. R. Feinstein and D. V. Cicchetti. 1990. High agreement but low kappa: I. the problem of two paradoxes. *Journal of Clinical Epidemiology*, 43:543–549.

Kilem Gwet. 2001. *Handbook of Inter-rater Reliability*. STATAXIS Publishing Company.

Klaus Krippendorff. 1980. *Content Analysis*, volume 5. SAGE Publications, Beverly Hills / London.

Peter Kühnlein and Jens Stegmann. 2003. Empirical Issues in Deictic Gestures. Technical Report 2003/03, Bielefeld University.

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

Jan-Torsten Milde and Ulrike Gut. 2001. The TASX-environment. In *Proceedings of the IRCS Workshop on linguistic databases, Philadelphia*.

Charles Sanders Peirce. 1965. *Collected Papers*, volume II. Harvard University Press, Cambridge, MA.

Martin J. Pickering and Simon Garrod. in press. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*.

Willard Van Orman Quine. 1960. *Word and Object*. M.I.T Press, Cambridge, MA.

R Development Core Team, 2003. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Hannes Rieser. 2004. Pointing in dialogue. In *CATALOG04 Conference Proceedings*. pp.

Ludwig Wittgenstein. 1958. *The Blue and Brown Books*. Harper & Row, New York.