Speech and 2D Deictic Gesture Reference to Virtual Scenes

Niels Ole Bernsen

Natural Interactive Systems Laboratory, Campusvej 55, DK-5230 Odense, Denmark
nob@nis.sdu.dk, http://www.nis.sdu.dk

Abstract. Humans make ample use of deictic gesture and spoken reference in referring to perceived phenomena in the spatial environment, such as visible objects, sound sources, tactile objects, or even sources of smell and taste. Multimodal and natural interactive systems developers are beginning to face the challenges involved in making systems correctly interpret user input belonging to this general class of multimodal references. This paper addresses a first fragment of the general problem, i.e., spoken and/or 2D on-screen deictic gesture reference to graphics output scenes. The approach is to confront existing sketchy theory with new data and generalise the results to what may be a more comprehensive understanding of the problem.

1 Introduction

Speech and deictic (pointing, delimiting, etc.) gesture input is known as an excellent multimodal input combination for interacting with many different kinds of application. The reason is that, by itself, unimodal speech is often poor at providing unambiguous reference to spatial objects, unambiguously specifying spatial manipulations for the system to do, etc. [Bernsen 2002, cf. Bolt 1980]. With today's rapidly evolving multimodal technologies, we will soon be able to provide robust camera-captured 3D deictic gesture input into virtual reality scenes. Still, spontaneous *2D* deictic gesture and spontaneous spoken input addressing 3D virtual reality scenes remains the state of the art. This may be viewed as a good thing from the point of view of scientific methodology. It induces us to first develop applicable theory for the speech and 2D deictic gesture input case before attempting to generalise the theory two-fold, i.e., to the general case of handling spontaneous speech and spontaneous (i) *3D* deictic gesture input which, moreover, (ii) not only refers to visually perceived scenes but also to aurally perceived 3D sound sources, tactilely perceived objects, olfactorily perceived sources of smell, and gustatorily perceived sources of taste.

As for the background of this paper, the author and colleagues received a sense of the complexity of the problem when developing speech/gesture input fusion for a domain-oriented (or non-task-oriented) system enabling English user conversation with 3D-embodied fairytale author Hans Christian Andersen [Martin et al. 2006, to appear]. In the course of conversation, users may refer, using spontaneous speech and/or 2D tactile-screen deictic gesture, to objects in Andersen's study which he might tell stories about. Figure 1.1 shows Andersen in front of some seven such objects, i.e., the six pictures on the wall and his pen on the writing desk. In the absence of applicable theory for semantic-level speech/gesture input fusion, our approach to input fusion in the Andersen system was a series of cautious design decisions which favoured confidence in gesture input over confidence in spoken input. Whilst largely successful, it was clear to us from the start that those decisions do not scale to applications in which the spoken input contents are typically richer than being mostly a semantically and pragmatically redundant reflection of the gesture input contents.

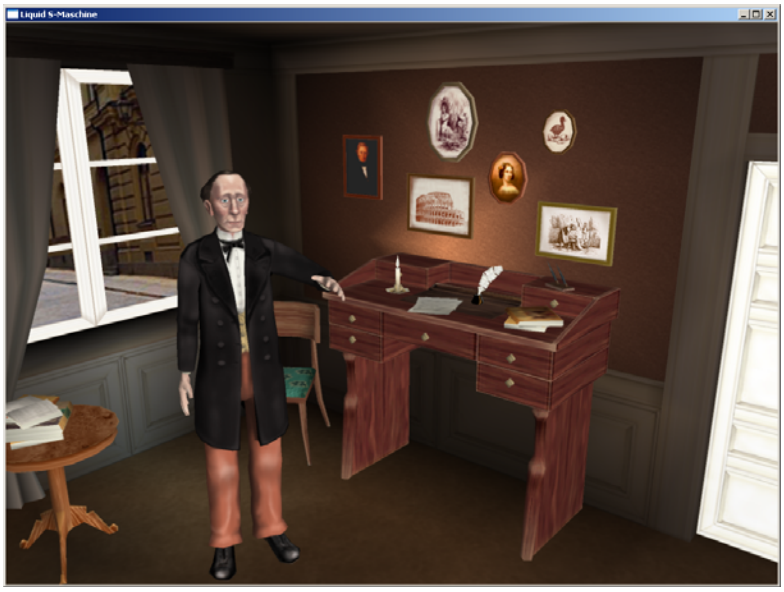

Figure 1.1. Hans Christian Andersen in his study surrounded by gesturable objects.

Even if not directly *applicable* to the speech/gesture problem in the Andersen system, there is, in fact, *relevant* theory around. In particular, [Landragin 2006, to appear] proposes a sketch of how to handle, theoretically as well as algorithmically, the speech and 2D tactile-screen deictic references to virtual objects. The paper is based primarily on a small corpus collected under controlled circumstances and on partial algorithmic implementation in two task-oriented system research projects. As Landragin points out, his solution sketch is potentially limited by the small and in other ways limited corpus he has had the opportunity to analyse.

The question to be addressed below can now be stated in simple terms. If we test Landragin's approach with data from a rather different corpus, what happens? Will the theory sketch stand or must it be significantly expanded to capture the full dimensionality of the problem? In the latter case, we will probably have to analyse additional corpora and do more conceptual homework before arriving at stable theory on which to base processing of spontaneous speech and 2D deictic gesture references. The data analysed and discussed in this paper is from the user test of the second Andersen system prototype.

2 The General Problem Approached

The general problem may be characterised as follows. (1) users may use speech and/or 2D deictic gesture (henceforth: gesture) to refer to virtual objects as part of dialogue or conversation. (2) The system must be able to interpret, and appropriately respond to, whichever referential input is provided by the users, no matter which task, domain, or interactive application the user is addressing. (3) We need a formal model of the problem that can reliably serve as a basis for algorithm development. It should be noted that the problem is not about speech/gesture input fusion *per se*, it is about spoken *and/or* 2D deictic gesture reference to virtual objects. Multimodal fusion is an issue only when the user speaks and gestures more or less simultaneously. Even in a universe of discourse in which this form of multimodal reference is possible, we often make unambiguous reference to scenes and objects using speech-only or gesture-only.

2.1 Landragin's Approach

A slightly extended version of Landragin's model is shown below. Referential speech/ gesture input may be represented as the quadruplet:

/ referring mode / grammatical category of referring expression / deictic gesture yes/no / other information / (**1**)

Referring mode is a pragmatic descriptor of the referential act made, such as *indicate a particular referent*. This descriptor is modality-independent in Landragin's model due to some tacit assumptions made. *Grammatical category* is the grammatical category of the spoken referring expression, such as *indefinite noun phrase*. *Deictic gesture yes/no* marks if the referential communicative act did or did not include deictic gesture. *Other information* has been added to Landragin's model. It enables us to add information required for input interpretation, such as that the input is anaphoric or elliptic. In addition, the model includes the notion of a *reference domain*, i.e., a subset of scene objects with something in common, such as forming a spatial group or being similar in shape, size, or perceptual salience. Humans introduce reference domains to simplify reference interpretation by sub-dividing the visible scene into sub-domains within which it is easier to disambiguate spoken or speech/gesture reference to particular entities. The nature of human perception must be taken into account because, e.g., the relative *salience* of visually perceived objects and the phenomenon of *perceptual grouping* are important underlying mechanisms for resolving referential ambiguity in conversation involving speech and 2D gesture.

The main corpus on which the model is based is a corpus of 98 data points from a Wizard of Oz exercise in which subjects had to manipulate, i.e., identify and request system actions onto, virtual objects, such as triangles and squares [Wolf et al. 1998].

Table 2.1 shows the result of Landragin's corpus analysis. Six *referring modes* were found. These referential actions are to: (1) introduce a new referent by creating it - *new-ref*; (2) extract any referent *-ext-any-ref-* or (3) extract a particular referent *-ext-par-ref-* from an already delimited reference domain; (4) indicate a particular referent that is, or has been, focused by, e.g., gesture or prior spoken reference -*ind-par-ref*; (5) indicate a particular reference domain, using gesture to focus on a particular object in the domain -*ind-par-dom*; (6) referring to a generic entity, e.g., 'triangles' -*gen-ref*.

The *grammatical categories* expressing the referring modes in the corpus are five: *indefinite noun phrases*, *definite noun phrases*, *demonstrative noun phrases*, *demonstrative pronouns*, and *personal pronouns*, see Table 2.1 for examples. The table also shows if deictic gesture accompanies spoken reference and adds other information.

Table 2.1. Landragin's corpus analysis. Dem is demonstrative, NP is noun phrase. P is pronoun

Referring mode	**Grammatical category**	**Gesture**	**Example**	**Other information**
new-ref	Indefinite NP	**no**	Create **a square**	Kataphor
ext-any-ref	Indefinite NP	**no**	Delete **a square**	Anaphor
ext-par-ref	Definite NP	yes	**The square**	
	Definite NP	**no**	**The square, The square** to the left	Anaphor or contextually obvious
	Dem NP	**no**	Delete **this square**	Anaphor
	Dem P	**no**	**This one**	Anaphor
ind-par-ref	Indefinite NP	yes	Delete **a square**	
	Definite NP	yes	**The square**	
	Definite NP	**no**	**The square**	Anaphor
	Dem NP	yes	**This square**	
	Dem NP	**no**	**This square**	Anaphor
	Personal P	**no**	Delete **it**	Anaphor
	Dem P	yes	**This one**	
ind-par-dom	Definite NP	yes	**The squares** [pointing to one of them]	
	Definite NP	**no**	**The group**	Anaphor
	Dem NP	yes	**These squares** [poin ting to one of them]	
	Dem NP	**no**	**This group**	Anaphor
	Personal P	**no**	Delete **them**	Anaphor
gen-ref	Indefinite NP	**no**	**A square** has four sides	
	Definite NP	**no**	**The square** has four sides	

	Dem NP	yes	**These forms** [pointing to one of them]	
	Dem NP	**no**	**These forms**	Anaphor
	Dem NP	yes	**This form**	
	Personal P	**no**	I have added a red square because **they** are eye-catching	Anaphor
	Dem P	**no**	**These ones** are eye-catching	Anaphor

2.2 Potential Need for Generalisation

How representative is the universe of referential discourse of Table 2.1 of the full complexity of the speech/2D gesture reference problem? Drawing upon Modality Theory [Bernsen 2002], the corpus has the following properties: (1) *static* graphics output domain in which the user perceives the *entire* collection of objects; (2) *mere output objects*, i.e., the geometric shapes in the corpus do not themselves represent information; (3) *2D objects* which do not allow for occlusion, objects resting on larger objects, etc.; (4) *simple and easy-to-label objects,* such as triangles and circles of different colour. In addition, (5) users' spoken input seems to be *simple and partially controlled* language. Users appear to speak explicitly at all times, always saying, e.g., "this triangle" rather than "this", i.e., using a demonstrative noun phrase when a pure demonstrative might suffice. What the users do in addition to object reference is to command simple changes, such as 'select', 'create' or 'delete'; (6) there are *no gesture-only object references*. Users never seem to shortcut by just pointing to objects.

Clearly, there are lots of applications which could benefit from multimodal speech/gesture input reference and which do not share the limitations just listed. It is thus a real possibility that, were we to analyse a corpus from an application domain with different general characteristics, we might find a different pattern of referential quadruplets, forcing extension to the quadruplet formal model *in spe*.

Let us use the method of generalisation-by-negation to broaden the scope of a general theory of speech/2D gesture reference, following the numbering above and introducing a couple of definitions (DEFn). *DEF1*: let us call what the user sees and what is being referred to, a visual or graphical *scene*. Note that the scene itself, and not just the objects *in* the scene, may have properties and be referred to. (1) The scene may be *static or dynamic*. It may include objects which move, change or act, and the scene itself may shift dynamically, e.g., because the user changes the virtual camera angle or the scene itself shifts beyond the user's control. *DEF2*: some scenes may be described as *scene worlds*, i.e., as the sum total of all possible scenes in the application, past, present and future. Thus, users might refer to past-but-not-present *scene world snapshots*, their objects and properties, and to future and expected but not-yet-perceived snapshots. (2) Scene objects may be either *mere objects*, such as a triangle, or *representational objects* which represent information, such as an image showing several objects. Users may correctly refer to this image using both singular and plural pronouns, as in “What is this [pointing]?” vs. “Who are they [pointing]?” [Martin et al. 2006, to appear]. (3) Scene worlds may be *2D or 3D*. 3D introduces new referentially relevant aspects, such as the backside of objects or objects viewed from above, occluded by other objects, or resting on larger objects. (4) Contrary to simple geometric shapes, *complex real-world-like objects* do not necessarily have a single standard label and are easier to mislabel when referring to them. (5) The general case of spoken input is uninstructed *spontaneous speech*. A general solution to our problem must be able to handle any kind of spoken reference to scenes, objects and properties, irrespective of linguistic complexity, speech acts performed, etc. (6) As for 2D deictic gesture, it must be assumed that users will sometimes make *gesture-only reference*. This can be done correctly and successfully, as we shall see, because application-specific deixis sometimes, at least, comes with implicit or explicit semantics and pragmatics.

There may be other dimensions along which we should generalise in order to describe the full scope of a theory of speech/2D gesture reference. However, the above generalisations demonstrate that the universe of referential speech-2D deictic gesture discourse is far larger than the one addressed in Landragin’s main corpus. In fact, those generalisations might provide an approximate target for a general theory.

Still, these are general arguments. We need to analyse new corpora in which users refer to scenes, objects and properties in sectors of the universe of referential speech-2D deictic gesture discourse other than those addressed by Landragin’s model in order to identify new specific types of referential discourse compared to those in Table 2.1.

3 A Different Corpus

In this section we present and analyse a corpus of English speech and 2D deictic gesture which represents a rather different fragment of the universe of referential discourse compared to Landragin’s main corpus. This new corpus reflects interaction with an application having all the properties generated in Section 2.2, including: a dynamic scene world; representational objects; 3D scenes and objects; complex photo-realistic, real-world-like objects with multiple properties; spontaneous conversational speech input; and gesture-only reference.

The corpus was produced as a follow-up to the user test of the second Hans Christian Andersen (HCA) system prototype made in February 2005. By contrast with the larger user test which involved Danish kids speaking English, the corpus to be discussed here was recorded with four native English speaking children, two girls and two boys, aged between 10 and 13 years. The native English test had two test conditions. In the first condition, the users were (i) instructed in how to use the keyboard for changing virtual camera angle and making HCA move when in non-autonomous mode, the tactile screen, and the microphone headset; and (ii) coached in spontaneous, free-style conversation with HCA, such as in re-phrasing input if not understood rather than just repeating the input. For the second condition which lasted 20-25 minutes per subject, the users were provided with a handout which proposed a series of

11 issues they could try to address in conversation at their leisure and in random order. We shall be looking at the second-condition corpus consisting of four conversations with HCA. All module interactions were logged and the spoken user input was recorded and transcribed. To correct for gesture recognition and interpretation errors relative to what the users actually did, the gesture log data was augmented with data from the two-camera video recordings of the user-system interactions. All HCA development and test corpora are available at NISLab's website [www.nis.sdu.dk] and are described more comprehensively in [Bernsen et al. 2006, to appear].

3.1 Corpus Annotation

With one major exception, all speech and/or 2D deictic gesture references to scene worlds and scenes, scene properties, and scene objects and their properties have been annotated. The exception is the many spoken input utterances in which the *sole* scene referent is HCA but in which no further reference is made to visual scene contents. Rather, reference is made to abstract discourse entities. Thus, e.g., an input utterance referring to "your hair" is annotated since HCA's hair is visible but, e.g., reference to "your fairytales" is not annotated because HCA's work is an abstract discourse object. Similarly, "you are ugly" is included since 'ugly' refers to visible properties of HCA, whereas, e.g., "you are cool" is not annotated. By implication, speech-only input which refers anaphorically to abstract discourse entities, such as the volunteered user comment "That is a sad story", is not annotated. If there is gesture, any accompanying speech is annotated for any reference.

Ignoring annotation beyond the scope of this paper, the corpus was annotated as follows: (1) Find the next data point to be annotated given the criteria stated at the start of this section. (2) If the data point includes spoken reference, identify the referring word or phrase and its grammatical category. If the category is in Landragin's coding scheme, mark this. If not, create new quadruplet. (3) Assign one of Landragin's referring modes to the spoken reference. Create new referring mode if referring mode is not in Landragin's coding scheme. (4) Mark if gesture accompanies spoken input. (5) Mark other relevant information.

The results of annotating the 78 data points in the corpus are shown in Table 3.1.

Table 3.1. Results of annotating the native English Andersen corpus. Dem is demonstrative, NP is noun phrase. P is pronoun, ref is reference, S is speech

Referring mode	**Grammatical category**	**Gesture**	**Example**	**Other information**
1. ind-par-ref	Dem NP	yes	Have you written **these books**?	
2.	Pure Dem	yes	What is **this**? **This**?	S redundant
3. gen-ref	Indefinite NP	yes	Did you always write with **a feather**?	
4.	Indefinite NP	yes	Do you like **trains** [point-	Indirect ref

			ing to picture of locomotive]	
5.	Indefinite NP	yes	Do you **read** a lot? [pointing to stack of books]	Indirect ref, ellipsis
6. scene-world-ref	Pure Dem	yes	Is **this** were you live? [circling gesture]	Metonym
7.	Pure Dem	**no**	Is **this** where you live?	Metonym
8. unique ref	Definite NP	**no**	**Your nose** is very big	Exophor
9.	Personal P	**no**	**You** are very old	Exophor
10. gesture-only-ref	N/A	yes	[pointing to anything]	N/A
11. indep-S-gesture-ref	Definite NP	yes	I do not like **your hair** [pointing to featherpen]	Exophor
12.	Personal P	yes	Have **you** been baptized? [pointing to featherpen]	Exophor

3.2 Annotation Analysis

Table 3.1 shows 12 quadruplets, only one of which, i.e., Quadruplet 1 */ ind-par-ref / Dem NP / yes / - /* is also present in Landragin's main corpus, cf. Table 2.1. Let us look at the 11 others in order, using Qn for Quadruplet (n).

Q2, the pure demonstrative *this* or, rarely, *that*, used in combined speech-gesture reference, occurred more frequently than any other quadruplet in the corpus with 38 or close to 50% of all data points. Q2 is always part of an explicit or implicit interrogative, such as "What is this?" We consider "What is this?" strictly *redundant* relative to the accompanying gesture, so let's explain the claim made earlier that deictic gesture itself can have a particular semantic and pragmatic meaning which may be relative to the application. HCA encourages the user to point to objects which he can tell stories about. So, when this happens, the meaning of the pointing *includes* the meaning of the spoken question "What is this?" and its variations in number or in elliptical expression. Note that the pointing gesture has its own pragmatic meaning in addition to the abstract spoken meaning pattern it includes. If the object referred to is not visually salient or in other ways prominent in the scene, the spoken "What is this?" may not succeed in uniquely referring to the object, whereas well-formed pointing to the object will always do that. So, the redundancy is asymmetric.

Q3, Q4 and Q5 are all *gen-ref* quadruplets in which the NP refers to *a kind* of object rather than to the particular object referred to in the accompanying gesture. From the point of view of input interpretation, Q4 and Q5 are of particular interest. In Q4 the user points to a picture, i.e., an object which itself represents information, of a *locomotive* but asks about *trains*. We call this *indirect reference* and the system must figure out, somehow, that the user is right in talking about trains here rather than declaring the user wrong or asking what the user means. In Q5 we also have a form of indirect reference, i.e., the ellipsis “Do you read [books] a lot?” or “Do you read a lot [of books]?” The system must figure this out.

Q6 and Q7 belong to our first new referring mode, i.e., *scene-world-ref*. In Q6, the user makes a circling gesture on the screen and asks about the entire scene world. What the user actually sees on the screen at the time is part of HCA’s study but the user refers to the study as a whole or possibly to HCA’s entire apartment. Since the question is not understood by HCA, the user repeats the question without accompanying gesture (Q7) and gets the correct response from HCA. We have marked the user’s pure demonstrative reference and encircling gesture as *metonymic* because the communicative intent is to refer to the whole by referring to part of it.

Q8 and Q9 belong to a second new referring mode, i.e., *unique-ref*. Without using gesture, the user succeeds in uniquely referring to on-screen objects using a definite NP and a personal pronoun, respectively. In reference theory, these grammatical forms would normally be anaphoric references which presuppose that their referents, i.e., HCA’s nose and his physical person, respectively, have been introduced already. This is why there are so many anaphoric quadruplets in Landragin’s main corpus, cf. Table 2.1. However, explicit prior referent introduction is not necessary in *exophoric* reference by which we successfully refer to prominent, or otherwise unique, objects, perceptually or otherwise, with no prior introduction. For instance, we don’t need to introduce “the sun” as referent prior to saying, e.g., “The sun is hot today”. In fact, Table 2.1 illustrates another reference phenomenon in addition the anaphor, i.e., *kataphoric* reference, in which we “refer ahead” to a referent that will be introduced later, as is, indeed, the case for an object which will only come into existence as an effect of the user’s command “Create a square”.

Q10 is a third new, *gesture-only-ref* referring mode which shows, furthermore, that not all referring modes are modality-independent (cf. Section 2.1). Q10 works perfectly well in the application context, we argue, because gesture already has a well-defined semantic-pragmatic meaning, i.e., what we would render linguistically as, e.g., “What is this?” This is why the system has no problem interpreting the user’s intended meaning and also why, had the user said at the same time “What is this?”, this spoken utterance would have been redundant relative to the deictic gesture (cf. above).

Finally, the fourth new referring mode, *indep-speech-gesture-ref*, highlights another limitation of Landragin’s corpus, i.e., that speech and gesture are always complementary in that corpus. Both *complementarity*, i.e., that speech and gesture both contribute necessary and non-redundant parts of the user’s intended meaning, and redundancy (cf. above), imply that speech and gesture are *semantically related and consistent*. In real life speech/gesture interaction, however, speech-gesture *inconsistency* is bound to occur from time to time and so is speech and gesture which is not semantically related but, rather, *independent* from each other. Although our small native English corpus does not have a case of the former phenomenon, Q11 and Q12 are cases of the latter.

4 Corpus Comparison

Section 2.2 presented a series of conceptual arguments why the interactive task and the application domain with which Landragin's main corpus was generated must be considered severely limited in many respects compared to the full potential universe of application of speech and 2D gesture. In the course of the argument, we introduced a series of notions of aspects of potential scenes which might be addressed through speech and 2D deictic gesture and which did not (appear to) apply to the scene and the interaction with it in the application used for collecting Landragin's main corpus. The notions were: *dynamic scenes, scene worlds, representational scene objects, 3D scenes and objects, complex, real-world-like objects, rich spontaneous speech,* and *gesture-only reference*. The implication was the hypothesis that a speech/2D deictic gesture corpus collected with a different form of interaction and a different application domain, might include quadruplets different from those listed in Table 2.1.

The scene world of the HCA system is characterised by all the notions introduced in Section 2.2. And the analysis of the native English HCA corpus in Section 3.2 confirms the hypothesis that a corpus of this nature would exhibit very different referential phenomena from those reported in Table 2.1. In fact, the confirmation is massive to the extent that only a *single* quadruplet among the 25 quadruplets in Table 2.1 was found in the HCA native English corpus. Even if [Landragin 2006, to appear] were right in claiming that this one, i.e., Quadruplet 1 in Table 3.1, */ ind-par-ref / Dem NP / yes / - /*, is the most common way of referring to scene objects using speech and 2D deictic gesture – the HCA data does not bear this out since pure demonstrative reference was done five times more frequently than reference using a demonstrative NP - it seems highly likely that we still have more to learn about the varieties of speech and 2D deictic gesture references. It would seem näive to claim that the *union* of Tables 2.1 and 3.1 comes close to constituting the total number of speech and 2D gesture reference phenomena. Rather, the user-HCA conversation provides a first taste of unconstrained speech/2D deictic gesture interaction and that's it. To better understand why, let us re-visit the corpus analyses in Tables 2.1 and 3.1.

4.1 Absent from HCA conversation but present in Landragin

In comparing the corpus analyses, we propose to focus on the quadruplet values *referring mode*, *gesture yes/no* and *other information*, considering *grammatical category* a more detailed quadruplet value to be worked out when we have a better grasp of the reference problem as a whole.

Among the six referring modes in Table 2.1, four were not found in the native English HCA corpus, i.e., *new-ref, ext-any-ref, ext-par-ref,* and *ind-par-dom* (see Section 2.1 for definitions). Among these, *new-ref* (e.g., "Create a square") and *ext-any-ref* (e.g., "Delete a square"), might be viewed as tied to a particular family of application. Also *ind-par-dom* in which a particular reference domain is indicated through a gesture that focuses on a particular object in the domain, might be in this category given the many-look-alike-objects-character of Landragin's main application domains. The absence of *ext-par-ref* which is used to extract a particular reference from a reference domain activated through gesture, verbal reference, description, previous references to objects in the domain, or visual salience – is more surprising.

However, it is easy to imagine slight extensions to the HCA system which would enable the occurrence of *new-ref* and *ext-any-ref*. All we need is an HCA who can act in certain ways when encouraged by the user. For *ind-par-dom* and *ext-par-ref*, we do not even need system modification. All we need for *ind-par-dom* to occur is a user who says, e.g., "Tell me about the pictures" [pointing to a single picture in Figure 1.1]. This just did not happen in our data. Similarly, all we need for *ext-par-ref* is a user who says, e.g., "Let us talk about the pictures above your desk" followed by "tell me about the one on the left". It may be concluded that the absence from the HCA corpus of some of the referring modes in Table 2.1 in no way implies that those referring modes are strongly tied to a particular family of applications.

A note on the scope of a theory of speech/2D deictic gesture reference concerns the Table 2.1 *gen-ref* cases of using an indefinite or definite noun phrase without accompanying gesture in descriptions of triangles-in-general. It is not obvious that these cases refer to scene aspects at all even though the scene objects include, e.g., triangles.

4.2 Present in HCA conversation but absent in Landragin

Section 3.2 describes four new referring modes, i.e., *scene-world-ref, unique-ref, gesture-only ref*, and *indep-speech-gesture-ref*. Of these, *scene world ref* enables reference to the scene world as a whole. Since the scene world is by definition not visible as a whole at any one time, pure demonstrative reference and deictic gesture-only reference to it must necessarily be metonymic whereas non-metonymic linguistic reference can be easily done. The referring mode *unique-ref* is exophoric reference to unique scene objects and properties. Exophoric reference does *not* depend on visual salience as discussed by Landragin but may be done to non-salient objects as long as these are unique in the scene world. *Gesture-only-ref* represents a much-needed theory extension acknowledging gesture-only reference. Finally, *indep-speech-gesture-ref* reflects another necessary theory extension which takes into account that more or less simultaneous input speech and deictic gesture may be semantically and pragmatically independent. In addition, we found some new *gen-ref* quadruplets, including some involving indirect reference and elliptical reference. Taking the two corpora together, we have found anaphoric, kataphoric and exophoric reference. And, having found speech-gesture complementarity, redundancy and independence, it is easy to predict that a new corpus could show speech-gesture inconsistency – an 11th referring mode.

5. Conclusion

If we make the common assumption that the sign of mature theory, such as a theory of speech and/or 2D deictic gesture reference to scene aspects, is its ability to predict and explain the large majority of data within its scope, the inevitable conclusion is that the data we have looked at is only the tip of the iceberg. Statistical convergence between the categories of the theory and the phenomena present in the data corpora used for its development would seem a long way off. We can forget about general algorithms for speech and/or 2D deictic gesture interpretation for the time being.

What is needed is, first of all, new speech and 2D deictic gesture corpora whose analysis can add more concepts for a general theory than those presented in Landragin and in this paper. Secondly, it is necessary to take a new look at the structure of a theory which could incorporate the results. A possible top-level organising principle is the distinction between speech-only reference, gesture-only reference and more or less simultaneous speech-gesture reference. We also need a more thorough analysis of the inventory of scene worlds than made above. We have had a glimpse of the fact that 2D deictic gesture reference is inherently complex, so, we should also look at, e.g., what is the significance of different deictic gesture shapes, such as points, circles, semi-circles, straight and curved lines, crosses involving several separate contacts with the screen, doodles, etc.; are there relevant differences between tactile-screen deictic 2D gesture and, e.g., mouse deictic gesture; what is the significance of the temporal aspects of speech/deictic gesture reference; and what is the semantics and pragmatics of 2D deictic gesture for different application families? We also need, at some point, a thorough theoretical comparison with, and possible multimodal extension of, linguistic reference theory [Kamp and Reyle 1993], or rather, perhaps, subsumption of both under a unified theory of multimodal reference, considering purely linguistic reference as a special case.

Finally, we need to take a system development point of view of it all. Like humans, the system must evaluate the input before planning its response. For instance, it must evaluate the truth of a definite NP like “This triangle [pointing]”. What if “this” object which is being both pointed to and classified as a triangle is *not* a triangle?

Acknowledgement and References

The HCA system was built in the NICE (Natural Interactive Conversation for Edutainment) project supported by the EU’s Human Language Technologies programme. The support is gratefully acknowledged. I would like to thank Svend Kiilerich for his video-based annotation correction of the HCA native English corpus.

1. Bernsen, N. O.: Multimodality in Language and Speech Systems - from Theory to Design Support Tool. In: Granström, B., House, D., and Karlsson, I. (eds.): Multimodality in Language and Speech Systems. Kluwer Academic Publishers, Dordrecht (2002) 93-148
2. Bernsen, N.O., Dybkjær, L., Kiilerich, S.: H.C. Andersen Conversation Corpus. In: Proceedings of the Language Resources and Evaluation Conference, Genoa, May 2006 (to appear)
3. Bolt, R.A.: Put-That-There: Voice and Gesture at the Graphics Interface. Computer Graphics 14(3) (1980) 262-270
4. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer Academic Publishers, Dordrecht (1993)
5. Landragin, F.: Visual Perception, Language and Gesture: A Model for their Understanding in Multimodal Dialogue Systems. Special Issue of on Multimodal Interaction, Signal Processing (2006, to appear)

6. Martin, J.C., Buisine, S., Pitel, G., Bernsen, N. O.: Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters. Special Issue of on Multimodal Interaction, Signal Processing (2006, to appear)
7. Wolf, A., De Angeli, Romary, L.: Acting on a Visual World: The Role of Perception in Multimodal HCI. In: Proceedings of the AAAI'98 Workshop: Representations for Multi-Modal Human-Computer Interaction. Madison, Wisconsin (1998)