

Holsanova (forthc.), in: Sachs-Hombach, K. & Totzke, R. (eds.) Bilder, Sehen, Denken. Herbert von Halem Verlag: Köln.

JANA HOLSANOVA

How we focus attention in picture viewing, picture description and mental imagery

1. Introduction

In his book *Visual thinking*, Arnheim draws the conclusion that »visual perception [...] is not a passive recording of stimulus material but an active concern of the mind« (Arnheim 1969: 37). From a cognitive science perspective, it is a major concern to study these processes empirically, in detail. We cannot *directly* uncover the contents of our mind but we can come closer to cognitive processes via overt manifestations. I am interested in how speakers perceive, conceptualise and spontaneously describe complex pictures on higher levels of discourse and I use eye tracking methodology along with verbal protocols and a multimodal scoring method to study these mental processes.

In this paper, I present results from my empirical research on picture viewing and picture description. The aim of the paper is to show how overt verbal and visual protocols can, in concert, elucidate covert mental processes. Section 1 focuses on the method for synchronizing verbal and visual data as two windows to the mind. In section 2, I show how viewers verbalize their impressions from pictures, how they categorize picture elements, interpret picture content and how their picture descriptions (and perceptions) differ in style. In section 3, I present our eye tracking studies on picture viewing and mental imagery. Finally, in section 4, I summarize and consider how the results from these studies contribute to an understanding of the dynamics of the ongoing perception processes – an area on the intersection of artistic and cognitive theories.

2. Picture viewing and picture description: Two windows on the mind

Eye movements have always been of great interest to cognitive scientists, since conscious control is related to human volition, intention and attention (Solso 1994). Empirical evidence suggests that the direction of eye movements and the direction of attention are linked (Sneider/Deubel 1995; Theeuwes 1993; Juola et al. 1991). Although covert attention has been discussed (Posner 1980), its role in normal visual scanning has been minimised (Findlay/Walker 1999). Eye movements have thus been used to obtain valid measurements of a person's interests and cognitive processes, in particular when watching works of art (Buswell 1935; Yarbus 1967). Just and Carpenter (1976, 1980) propose a strong eye-mind assumption: when looking at relevant visual displays, there is no lag between what the observer is fixating and what the mind is processing. In contrast, Underwood and Everatt (1992) argue that one current fixation can indicate either past, present and future information acquisition, which weakens the strong eye-mind assumption. Eye fixations have been considered a boundary between perception and cognition, since they overtly indicate that information was acquired. Thanks to new technology, eye movements »provide an unobtrusive, sensitive, real-time behavioural index of ongoing visual and cognitive processing« (Henderson/Ferreira 2004: 18).

Another way of approaching the mind is via *spoken language*. Within psycholinguistic research, the spurt-like character of speech and dysfluencies in spoken language production have been explored in order to reveal planning and monitoring activities on different levels of discourse (Garett 1980). The hypothesis behind these approaches is that dysfluencies that belong to the flow of language production reflect the flow of thought. In psychology and cognitive science, spoken language has been used to externalise mental processes during different tasks in form of verbal protocols or think-aloud protocols (Ericsson/Simon 1980) where subjects are asked to verbally report sequences of thought during different tasks. Verbal protocols have been extensively used to reveal steps in reasoning, decision-making and problem-solving processes and applied as a tool for design and usability testing. Linguists were trying to access the mind through spoken descriptive discourse (Linde/Labov 1975) and through a consciousness-based analysis of narrative discourse (Chafe 1996).

In my studies, I combine the analysis of data from two sources, *eye movement data* from picture viewing and simultaneous spoken *picture descriptions*, with the expectation that these two kinds of data can give us distinct hints about the dynamics

of the underlying cognitive processes. On the one hand, eye movements reflect human thought processes. It is easy to determine which elements attract the observer's gaze, in what order and how often. In short, eye movements offer us a window to the mind. On the other hand, verbal foci formulated during picture description are the linguistic expressions of a conscious focus of attention. With the help of a segmented transcript, we can learn what is in the verbal focus at a given time. In short, spoken language description offers us another window to the mind. Both kinds of data are used as an indirect source to gain insights about the underlying cognitive processes and about human information processing. The way to uncover the ›idea unit‹ goes via spoken language in action and via the process of visual focusing.

2.1 *Focusing attention: The spotlight metaphor*

The human ability to focus attention on a smaller part of the visual field has been discussed in the literature and likened to a spotlight (Posner 1980; Olshausen/Koch 1995) or to a zoom-lens (Findlay/Walker 1998).

Actually, the ›spotlight‹ comes from the inner world, from the mind. It is the human ability to visually and verbally focus attention on one part of the information flow at a time. Here, this spotlight is transformed to both a verbal and a visual beam (cf. figure 1).

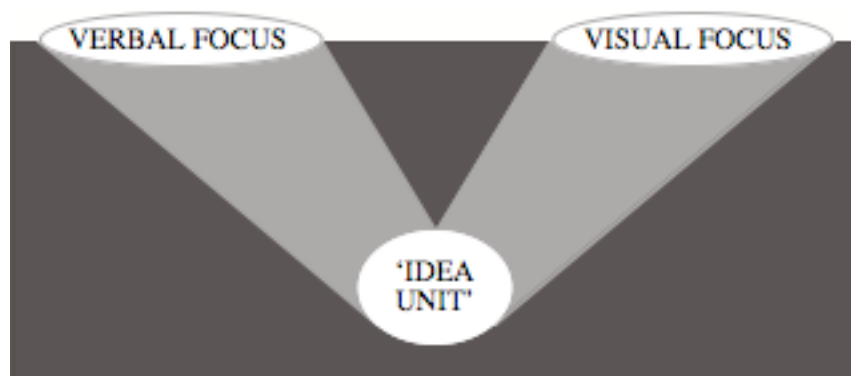


Fig. 1: Verbal focus, visual focus and ›idea unit‹ (Holsanova 2008:8)

What we attend to during visual perception and spoken language description can be conceived of with the help of a spotlight metaphor which intuitively provides a notion of limitation and focus. The picture elements fixated are visually in the focus of a spotlight and embedded in a context. The spotlight moves to the next area that

pops up from the periphery and will be in focus for a while. If we focus our concentration and eye movements on a point, we mostly also divert our attention to that point. By using a sequential visual fixation analysis, we can follow the path of attention deployed by the observer. Concerning spoken language description, it has been shown that we focus on one idea at a time (Chafe 1980, 1994). The picture elements described are in the focus of an attentional spotlight and embedded in the discourse context.

2.2 *Multimodal scoring method*

In order to analyse the relationship between the content of the visual focus of attention (specifically clusters of visual fixations) and the content of the verbal focus of attention (specifically clusters of verbal foci), I created temporally ordered multimodal score sheets. A multimodal time-coded score sheet is a format suitable for synchronising and analysing visual and verbal data (for details see Holsanova 2001: 99f.). Let us first look at figure 2 showing the result of viewing ›three birds in the tree‹.



Fig. 2: Fixation pattern: ›Three birds in the tree‹.(Holsanova 2008:132)

The motif comes from a children's book by Sven Nordqvist (1990). The fixation pattern shows us the path of picture discovery, that is, what objects and areas were fixated by the viewer and for how long. This is, though, a static pattern since it does not exactly visualise when and in what order they were fixated. The circles indicate only the position and duration of the fixations, the diameter of each fixation being proportional to its duration. The lines connecting fixations represent saccades. The white circle in the lower right corner is a reference point: it represents the size of a

one-second fixation. Let us now look at the transcript from the description of ›three birds in the tree‹.

Example 1: ›three birds in the tree‹

0310	(2s) and in the <i>middle</i> of the field there is a <i>tree</i>
0311	with one (1s) with three birds
0312	that are doing different things
0313	(1s) one bird is sitting on its <i>eggs</i> in a <i>nest</i>
0314	(1s) and the other bird (LAUGHING) is <i>singing</i>
0315	at the same time as the <i>third</i> like <i>female bird</i>
0316	(1s) is beating the <i>rug</i> or something,

Each numbered line in the transcript represents a new verbal focus expressing the content of active consciousness. Verbal focus is usually a phrase or a short clause, delimited by prosodic and acoustic features: it has one primary accent, a coherent intonation contour, and is usually preceded by a pause or hesitation (Holsanova 2001: 15f.). It implies that one new idea is formulated at a time and active information is replaced by other, partially different information at approximately two-second intervals (Chafe 1994). Several verbal foci are clustered into superfoci (for example summarizing superfocus or a list of items in the above example, delimited by lines). Verbal superfocus is a coherent chunk of speech, typically a longer sentence, consisting of several foci connected by the same thematic aspect and having a sentence final prosodic pattern. Superfoci can be conceived of as thresholds into a new complex unit of thought.

Instead of analysing the *result* of the viewing of ›three birds in the tree‹ in form of a fixation pattern (Fig. 2) and the *result* of the picture description in form of a transcribed extract (Example 1), I was able to visualise and analyse the *process* of picture viewing and picture description on a multimodal time-coded score sheet (Fig. 3).

As shown in Figure 3, the score sheet contains two different streams: it shows visual behaviour (objects fixated visually during description on line 1; thin line = short fixation duration; thick box = long fixation) and verbal behaviour (verbal idea units on line 2), synchronised over time. Since we start from the descriptive discourse level, units of discourse are marked and correlated with the visual behaviour. Simple bars mark the borders of verbal foci (expressing the conscious focus of attention) and

double bars mark the borders of verbal superfoci (thematic clusters of foci that form more complex units of thought). On line 3, we find the coding of superfocus types (*summarising, substantive, evaluative* etc.).

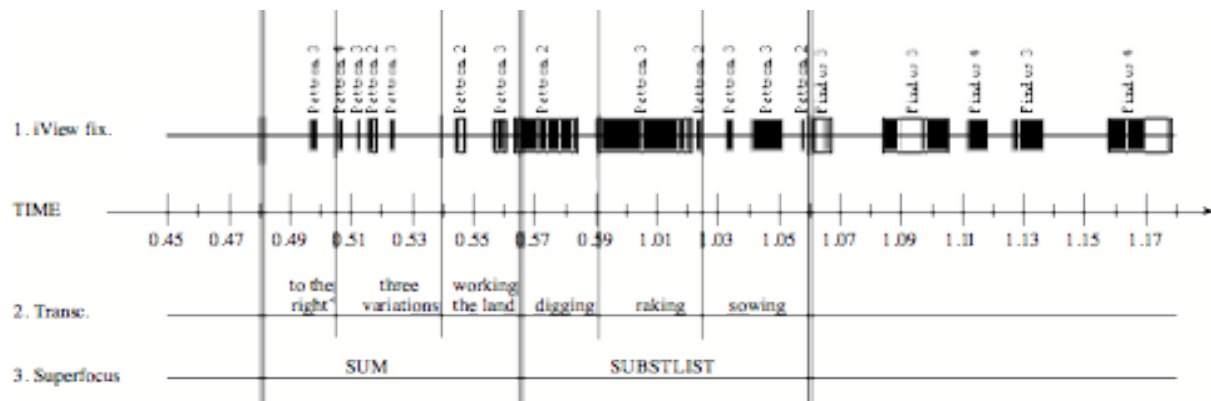


Fig. 3: Multimodal time-coded score sheet (Holsanova 2008:88)

With the help of this new analytic format, I could examine what was lying in the visual and verbal attentional spotlight at a particular moment: Configurations of verbal and visual clusters could be extracted and contents in the focused verbal idea flow and the visual fixation clusters could be compared.

This section gave an introduction to the method and analytic tools used in a series of empirical studies on picture viewing and picture description. By combining spoken language description and eye movement protocols, we can get insights into what the observers found interesting, what drew their attention, and how the scene was perceived. The multimodal score sheet enables us to synchronize visual and verbal behavior over time, to follow and compare the content of the attentional spotlight and to extract clusters in the visual and verbal flow.

3. *Understanding the dynamics of the ongoing perception processes*

With the help of the multimodal method described above, we will now take a closer look at the ongoing perception process. Thanks to the combination of verbal and visual data, mental processes and attitudes can be illuminated. In section 3.1, we will follow the viewers' categorizing and interpreting activities. In section 3.2, we will witness the process of spatial, semantic and mental grouping.

3.1 *Categorizing and interpreting*

When looking at a picture and describing it verbally, viewers are not only reporting about WHAT they see but also about HOW the picture appears to them. In other words, they are involved in both categorizing and interpreting activities:

As for the categorisation, the main figure, Pettson, can be described as a person, a man, an old guy, a farmer, or he can be called by his name. His appearance can be described (*a weird guy*), his clothes (*wearing a hat*), the activity he is involved in can be specified (*he is digging*) and this activity can be evaluated (*he is digging frenetically*). Thus, we can envision a description on different levels of specificity and with different degrees of creativity and freedom.

During evaluations, observers check the details (form, colour, contours, size) but also match the concrete, extracted features with the expected features of a prototypical/similar object: *›something that looks like a telephone line or telephone poles that is far too little in relation to the humans‹*. They compare objects both inside the picture world and outside of it. By using metaphors from other domains they compare the animal world with the human one: *›one bird is looking very human‹*; *›there's a dragonfly like a double aeroplane‹*.

Concerning introspection and report on mental states, let us look at the following example where one informant, after one minute of picture viewing and picture description, verbalises the following idea: *›when I think about it then it seems as if it in fact were two different fields, one can interpret it as if they were in two different fields' those persons here.‹*

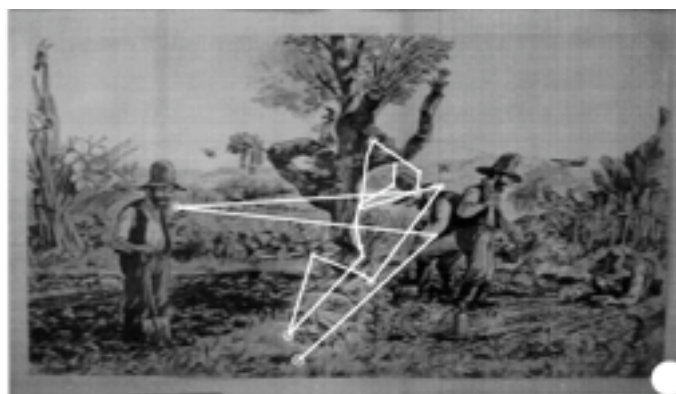


Fig. 4: Fixation pattern: *›Two different fields‹* (Holsanova 2008:133)

When inspecting the fixation pattern in figure 4, we can see that the observer is rescanning several earlier identified objects that are distributed in the scene. This is consistent with Yarbus' (1967) finding that eye movements occur in cycles and that

observers return to the same picture elements several times. This phenomenon has been confirmed by Noton and Stark (1971a, b) who coined the word ›scanpath‹ to describe the repetitive fixations in scene viewing. What we do not see and know, however, is how the observer perceives the objects on different occasions. I would claim that the objects that are refixated represent a bigger (compositional) portion of the picture and give support for his reconceptualization. By mentally zooming out, he discovers an inferential boundary between parts of the picture that he has not perceived before. The scene originally perceived in terms of ›one field‹ has become ›two fields‹ as the observer gets more and more acquainted with the picture. This example illustrates the dynamics of that the observer’s perception of the picture unfolds dynamically and can change over time.

It is also interesting to follow the perception and description of *repetitive figures* in the picture. The chosen children’s book illustration is a so-called *Simultanbild* containing multiple simultaneous persons and cats. *Simultanbild* is depicting a number of events that are spatially and temporarily not connected and strengthening the narrative mode of the picture by suggesting a continuation of an event. Kalkofen, Körber and Strack (this volume) who empirically studied perception of repetitive pictures found that multiple simultaneous persons in the picture are not noticed by the viewers. The authors suggest that this is due to ›repetition blindness‹. In our eye tracking studies on picture viewing and picture description, some viewers noticed repetitive figures (as documented in example 2) whereas other viewers did not mention them. In example 3, the viewer is clearly focusing on an event and describing one person and one cat involved in various activities.

Example 2:

0302	I see three/ eh <i>four</i> persons
0303	who are <i>digging</i> in a (2s) in a <i>field</i>
0304	or one person is <i>digging</i>
0305	(1s) another person is <i>raking</i>
0306	(1s) a third person is standing and <i>looking</i> at something
0307	at the same time as he is holding a <i>spade</i>
0308	and a fourth person seems to <i>sow things</i> ,
0309	(2s) in this field,

Example 3:

1102	it's a picture about <i>sowing</i>
1103	and it's a <i>fellow</i>
1104	I don't remember his name
1105	<i>Pettson</i> I think,
1106	who is first <i>digging</i> in the <i>field</i>
1107	and then he's <i>sowing</i>
1108	quite big white <i>seeds</i>
1109	and then the <i>cat</i> starts <i>watering</i>
1110	these particular <i>seeds</i>
1111	and it ends up with the fellow eh <i>raking</i> , this field,

The reason why some viewers mention repetitive figures and others do not could be due to the fact that some of them see the picture primarily as a representation whereas others focus on the content of the represented scene (cf. our discussion of the duality of picture perception in connection to description styles in section 3). There might as well be different ›perception styles‹.

Finally, let us mention the overall tendencies that could be extracted from the data on picture viewing and picture description. Informants start either with a specific categorisation that is then modified (*in a potato field or something; the second bird is screaming or something like that*) or, more often, a vague categorisation ›filler‹ is followed by a specification: *Pettson has found something he is looking at, he is looking at the soil*. In other words, a general characteristic of an object is successively exchanged against more specified guesses: *he is standing and looking at something maybe a stone he has in his hand that he has dug out*. When extracting an activity, for example when formulating ›Pettson is digging‹, the eye movements depict the relationships among picture elements and mimic it in repetitive rhythmic units, by filling in links between functionally close parts of an object (face - hand - tool; hand - tool, hand - tool - soil). In terms of cognitive semantics we can say that eye movements fill in the relationship according to a schema for an action.

In sum, the combination of visual and verbal data shows that the objects are focused on and conceptualised on different levels of specificity, objects' location and attributes are described and evaluated, judgements about properties and relations between picture elements are formulated, metaphors are used as a means of comparison and finally, the object's activity is described. What we have witnessed

here is a process of stepwise specification, evaluation, interpretation and even reconceptualisation of picture elements and the picture as a whole.

3.2 *Spatial, semantic and mental grouping*

According to Gestalt psychology, the organizing principles which enable us to perceive the patterns of stimuli as meaningful wholes are defined as (a) proximity, (b) similarity, (c) closure, (d) continuation, and (e) symmetry. The *proximity* principle implies that objects placed close to each other appear as groups rather than a random cluster. The *similarity* principle means that there is a tendency for elements of the same shape or colour to be seen as belonging together. Finally, the *symmetry* principle means that regions bounded by symmetrical borders tend to be perceived as coherent figures.

We can also think of several other principles of grouping that could have guided the informants' picture description: categorical or taxonomical proximity (elements that are similar to each other are described together), compositional principle (units delimited by composition are described together), saliency principle (expected, preferred and important elements are described first), animacy principle (human beings and other animate elements are described first) or even other principles.

The cluster *three birds* that we saw in figure 3 is based on the picture-inherent spatial proximity of objects. It is claimed that observers tend to fixate those closest objects first. However, observers also fixated and described clusters that were *not* spatially close. One type of cluster was based on groupings of multiple similar concrete objects (*four cats, four versions of Pettson*). This type of cluster could not be based on spatial proximity because the picture elements are distributed over the whole scene. In this case, we could rather speak about categorial proximity. (The simultaneity of the objects involved in different activities has probably promoted this guidance.) This type of cluster is usually described in a summarising superfocus that is typically followed by a list where each instance is described in detail.

Another type of cluster that was perceived as a meaningful unit in the scene was *hills on the horizon*. The observers' eyes followed the horizontal line, filling in links between objects. This cluster was probably a compositionally guided cluster. The observer is zooming out, scanning picture elements on a compositional level.

We are now moving further away from the scene inherent spatial proximity and coming to the next type of clusters constructed by the observers. This time, the

cluster is an example of an active mental grouping of concrete objects based on an extraction of similar traits and activities (see figure 5 ›flying objects‹). The prerequisite for this kind of grouping is a high level of active processing. Despite the fact that the objects are distributed across the whole scene and appear quite differently, they are perceived of as a unit because of the identified common denominator. The observer is mentally zooming out and creating a unit relatively ›independent‹ of the ›suggested‹ meaningful units in the scene. The eye movements mimic the describers' functional grouping of objects.

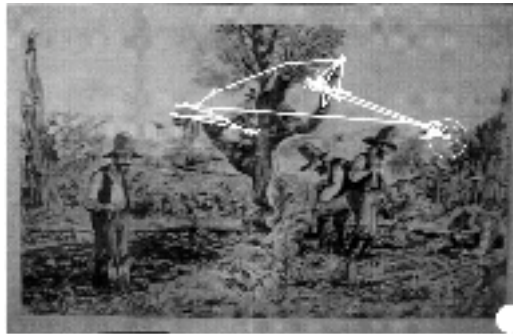


Fig. 5: Fixation patterns for ›Flying insects‹ (Holsanova 2008:134)

In a number of cases, especially in later parts of the observation, the clusters were based on thematic aspects and guided by top-down factors. The next cluster lacks spatial proximity (Figure 6). The observer is verbalising his impression about the picture content: ›it looks like early summer‹. This abstract concept is not identical with one or several concrete objects compositionally clustered in the scene, and visual fixations are not guided by spatial proximity. The previous scanning of the scene has led the observer to an indirect conclusion about the season of the year. In the visual fixation pattern, we can see large saccades across the whole picture composition. It is obviously a cluster based on mental coupling of concrete objects; their parts or attributes (such as flowers, foliage, plants, leaves, colours) are on a higher level of abstraction.

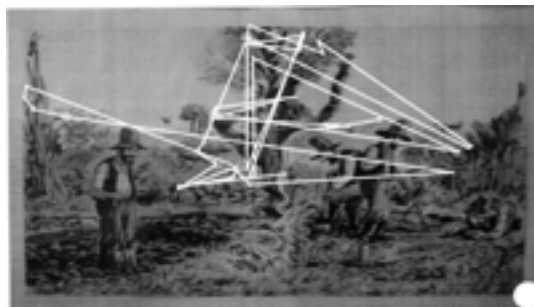


Fig. 6: Fixation patterns by a person saying ›It looks like in early summer‹. (Holsanova 2008:135)

Since the concept ›early summer‹ is not directly identical with one or several concrete objects in the scene, the semantic relation cannot be a mere referential one. Instead, the relation is inferred, the objects being concrete indices of a complex (abstract) concept. In addition, the relation between the spoken description and the visual depiction is not a categorical one, associated with object identification. Instead, the observer is formulating how the picture appears to him, on an abstract level. Afterwards, during visual rescanning, the observer is searching again for concrete objects and their parts as crucial indicators for his abstract statement. By refocusing these elements, the observer is in a way collecting evidence for his statement. In other words, he is checking whether the object characteristics in the concrete scene match with the symptoms for the described scenario. Concrete objects can be viewed differently on different occasions as a result of our mental zooming in and out. We have the ability to look at a single concrete object and simultaneously zoom out and speak about an abstract concept or about the picture as a whole. When talking about creativity and freedom, this type of ›mental groupings‹ shows a high degree of creativity.

In sum, apart from scene inherent concrete picture elements with spatial proximity, even new groupings were created on the way. Objects across the scene – horizontally or vertically aligned – were grouped due to the scene composition. Multiple similar elements distributed in the scene were clustered on the basis of a common taxonomic category. Active mental groupings were created on the basis of similar traits and common activity. The process of mental zooming in and out could be documented, where concrete objects were refixated and viewed on another specificity level or with another concept in mind. During their successive picture discovery, the informants also created new ›mentak‹ groupings across the whole scene, based on abstract concepts.

3.3 Variations in picture description: two different styles

After a qualitative analysis of picture descriptions where participants describe the picture from memory, I could extract two different styles of description, deploying different perspectives. Attending to spatial relations was dominant in the *static*

description style, while attending to the flow of time was the dominant pattern in the *dynamic description style*. Consider the following two examples:

Example 4:

0601	well it's a <i>picture</i>
0602	<i>rectangular</i>
0603	and ... it was mainly <i>green</i>
0604	with a blue <i>sky</i> ,
0605	divided into <i>foreground</i> and <i>background</i> ,
0606	in the middle there's a <i>tree</i>
0607	uh and in it three <i>birds</i> are sitting,
0608	the lowest bird on the left sits on her <i>eggs</i>
0609	and <i>above</i> her
0610	there is a bigger bird <i>standing up</i> ,

Example 5:

0702	it's quite early in the <i>spring</i>
0703	and Pettson and his cat Findus are going to <i>sow</i> ,
0704	Pettson starts with digging up his little <i>garden</i>
0705	then he <i>rakes</i>
0706	and then he sows <i>plants</i>
0707	uh puts <i>potatoes</i> out later on
0708	and when he's <i>ready</i>
0709	he starts sowing <i>lettuce</i> ,

Extract (4) is a prototypical example of a static description. In what could be called a technical style, the picture is typically decomposed into fields that are then described systematically, using a variety of terms for spatial relations. In the course of description, informants establish an elaborate set of referential frames that are used for localisations. They give a precise number of picture elements, stating their colour, geometric form and position. Apart from spatial expressions, the typical features of the technical description style are frequent use of nouns, existential constructions (>*there is*<, >*it is*<, >*it was*<), auxiliary verbs and passive voice.

In the dynamic style (as in extract 5), observers primarily focus on temporal relations and dynamic events in the picture. Although there is no temporal or causal order inherent in the picture, viewers infer it: They explicitly mark that they are talking about steps in a process, about successive phases, about a certain order. The

dynamic quality of this style is achieved by a frequent use of temporal verbs, temporal adverbs and motion verbs in active voice. Discourse markers are often used to focus and refocus on the picture elements, and to interconnect them. Apart from the above mentioned features, the informants seem to follow a narrative schema: the descriptions start with an introduction of the main characters, their involvement in various activities, and a description of the scene.

What are then the possible explanations for these two styles found in the picture descriptions? One observation concerns gender differences since we could see a preliminary tendency for women to be more dynamic and for men to have a more static description style. However, more studies are needed to confirm this pattern. Another possibility would be the difference between ›visual thinkers‹ and ›verbal thinkers‹ discussed in Holsanova (2008).

Furthermore, an additional source of explanation are the cognitive, experiential and contextual factors that might to some extent have influenced what people focus on during picture viewing, what they remember afterwards and what they describe verbally (and how). For instance, previous knowledge of the picture, of the genre, of the book, of the characters or of the story may be the most critical factors for the distinction of the two description styles, since not knowing the genre and / or characters may lead to a less dynamic description. One could expect that if the informants have read the book (to themselves or to their children), the picture will remind them of the story and the style will become dynamic and narrative. If the informants know the Pettson and Findus characters from Sven Nordqvist's books, films, TV programmes, computer games, calendars etc., they can easily identify the activities that these characters are usually involved in, and the description will become dynamic. Finally, if the informants know the particular story from the children's book, they can go over to a ›story-telling-mode‹ and deliver a dynamic description with narrative elements.

Yet another possibility is that these description styles are picture specific – in contrast to other ›sources of perception‹. It could be the case that some of the results are due to the fact that the informants either verbalise the picture as a representation or focus on the content of the represented scene. This explanation can be traced back to the concept of duality of picture perception, often described as a challenge to current theories of visual perception. Duality of picture perception is based on the phenomenon that we both see a *picture plane*, i.e. a physical part of the perceived

surrounding world, and a (perceived) *pictorial space*, i.e. the world shown in the picture. In connection to the two demonstrated picture description styles, it could be argued that the static description style (as exemplified in extract 4) reflects the priority to perceive the picture plane, whereas the dynamic description style (as exemplified in extract 5) reflects the priority to perceive pictorial space and the associated temporal relations and dynamic events.

This section has been devoted to variations in picture descriptions. Two different tendencies have been observed: the preference to focus on spatial relations in the picture typical of the static technical description style and the preference to focus on dynamic events and temporal relations in the picture typical of the dynamic narrative style. We have considered various cognitive, experiential and contextual factors that might have played a role for the picture description style.

4. Picture viewing and mental imagery

Experiences of having mental images are apparent in a lot of everyday situations. We ›see‹ images when we mentally recreate experiences, when we plan future events, when we solve problems, when we retrieve information about physical properties or relationships, or when we read an absorbing novel. Finke describes imagery as »the mental invention of an experience that at least in some respects resembles the experience of actually perceiving an object or an event either in conjunction with, or in the absence of, direct sensory stimulation« (Finke 1989: 2). In popular terms, mental imagery is often referred to as ›seeing something in the mind's eye‹.

How can we observe this phenomenon and prove that we think in images? Eye tracking methodology has become a very important tool in the study of human cognition, and current research has found a close relation between eye movements and mental imagery. Already in our first eye tracking study (Holsanova et al. 1998), we found some striking similarities between the informants' eye movement patterns when looking at a complex picture and their eye movements when looking at a white board and describing the picture from memory. The results were interpreted as a bit of support for the hypothesis that mental scanning (Kosslyn 1980) is used as an aid in recalling picture elements, especially when describing their visual and spatial properties. Brandt and Stark (1997) and Laeng and Teodorescu (2001) have shown that spontaneous eye movements closely reflect the content and spatial relations

from the original picture or scene. In order to extend these findings, we conducted a number of new eye tracking studies, both with verbal and pictorial elicitation, and both in light and in darkness (Johansson/Holsanova/Holmqvist 2005, 2006).

In the following, I will mention one of these studies. The experiment consisted of two main phases, a *viewing* phase in which the participants inspected the complex picture and a *description* phase in which the participants with their own words described this picture from memory while looking at a white screen. Eye movements were recorded during both phases. The descriptions were transcribed in order to analyse *which* picture elements were mentioned and *when*. The eye movements were then analysed according to objects derived from the descriptions. For instance, when an informant formulated the following superfocus,

Example 5:

01:20 – And eh hh to the <i>left</i> in the picture
01:23 – there are large <i>daffodils</i> ,
01:26 – it looks like there were also some <i>animals there</i> perhaps,

we would expect the informant to move her eyes towards the left part of the white screen during the first focus. Then it would be plausible to inspect the referent of the second focus (the daffodils). Finally, we could expect the informant to dwell for some time within the daffodil area – on the white screen – searching for the animals (three birds, in fact) that were sitting there on the picture.

The following criteria were applied in the analysis in order to judge whether correct eye movements occurred: Eye movements were considered correct in *local correspondence* when they moved from one position to another in the correct direction within a certain time interval. Eye movements were considered correct in *global correspondence* when moving from one position to another and finishing in a position that was spatially correct relative to the whole eye-tracking pattern of the informant (for a detailed description of our method, cf. Johansson et al. 2005, 2006). We tested significance between the number of correct eye movements and the expected number of correct movements by chance.

Our results were significant both in the local correspondence coding (74.8 percent correct eye movements, $p = 0,0015$) and in the global correspondence coding (54.9 percent correct eye movements, $p = 0,0051$). The results suggest that informants

visualise the spatial configuration of the scene as a support for their descriptions from memory. The effect we measured is robust. Our data indicate that even for a complex picture, spatial locations are to a high degree preserved when describing it from memory.

Resizing effects, i.e. the fact that subjects shrunk, enlarged and stretched the image, were quite common during picture description. It was also common that subjects re-centered the image from time to time; thus yielding local correspondence. Overall, there was a good similarity between data from the viewing and the description phases, as can be seen in Figure 7.



Fig. 7: Same subject: viewing phase (A) and description phase (B). (Johansson, Holsanova & Holmqvist 2006:1066).

Our results can be interpreted within different theoretical frameworks. According to Kosslyn (1994), distance, location and orientation of the mental image can be represented in the visual buffer, and it is possible to *shift attention* to certain parts or aspects of it. Laeng and Teodorescu (2001) interpret their results as a confirmation that eye movements play a functional role during image generation. Mast and Kosslyn (2002) propose that eye movements are stored as spatial indexes that are used to arrange the parts of the image correctly. Our results can be interpreted as further evidence that eye movements play a functional role in visual mental imagery and that eye movements indeed are stored as spatial indexes that are used to arrange the different parts correctly when a mental image is generated.

There are, however, alternative interpretations. Researchers within the 'embodied' view claim that instead of relying on a mental image, we use features in the external

environment. An imagined scene can then be projected over those external features, and any storing of the whole scene internally would be unnecessary. Ballard et al. (1996, 1997) suggest that informants leave behind ›deictic pointers‹ to locations of the scene in the environment, which later may be perceptually accessed when needed. Pylyshyn (2001) has developed a somewhat similar approach to support propositional representations and speaks about ›visual indices‹ (cf. also Spivey et al. 2004).

Yet another alternative explanation comes from ›simulation‹ accounts, where imagery experiences are not considered to rely on the format of image representations at all. Instead it is assumed that perception can be internally simulated by activating necessary regions in the brain (Hesslow 2002). The ›perceptual activity theory‹ suggests that instead of storing images, we store a continually updated and refined set of procedures or schemas that specify how to direct our attention in different situations (Thomas 1999). In this view, a perceptual experience consists of an ongoing, schema-guided perceptual exploration of the environment. Imagery is then the *re-enactment* of the specific exploratory perceptual behaviour that would be appropriate for exploring the imagined object as if it were actually present. Eye movements would thus occur when these re-enactments happen during mental imagery experiences. Barsalou (1999) suggests a similar approach to mental images in his ›perceptual symbol systems‹. A perceptual symbol is in this sense *not* a mental image, but a record of the neural activation that arises during perception. Imagery is then the re-enactment or simulation of the neural activity.

In this section, I introduced our study in picture viewing and mental imagery. A significant similarity was found between (a) the eye movement patterns during picture viewing and (b) those produced during picture description (when the subjects were looking at a white screen). The eye movements closely reflected the content and the spatial relations of the original picture suggesting that the subjects created some sort of mental image as an aid for their descriptions from memory.

5. Summary and discussion

Current theories in visual perception stress the cognitive basis of art and scene perception. We ›think‹ art as much as we ›see‹ art (Solso 1994). Our way of

perceiving objects in a scene can be triggered by our expectations, our interests, intentions, previous knowledge, context or instructions.

The aim of the paper was to show how overt verbal and visual protocols can, in concert, elucidate covert mental processes. This combination of data illuminates mental processes and attitudes and can thus be used as a sensitive tool for understanding the dynamics of the ongoing perception process. My data suggest that it is not only recognition of objects that matters but also how the picture appears to the viewers. Verbal descriptions include quality of experience, subjective content and mental state. Viewers report about (i) referents, states and events, (ii) colours, sizes and attributes, (iii) compositional aspects, (iv) they mentally group the perceived objects into more abstract entities, (v) compare picture elements, (vi) express attitudes, associations, (vii) report about mental states.

Subjects start by looking at picture-inherent objects, units and gestalts. As their picture viewing progresses, they tend to create mental units that are more independent of the concrete picture elements. They may make large saccades across the whole picture, picking up information from different locations to support concepts which are distributed across the picture (like ›early summer‹ or ›flying insects‹). With the increasing cognitive involvement, observers and describers tend to return to certain areas, change their perspective and reformulate or recategorise the scene. It becomes clear that perception of the picture changes over time. The dynamics of this categorization process is reflected in the usage of many refixations in picture viewing and reformulations, paraphrases and modifications in picture description. Finally, during description from memory, the subjects seem to ›redraw‹ or ›reconstruct‹ certain features from the original picture with the help of mental imagery.

In general, the results can contribute to the area on the intersection of artistic and cognitive theories. In particular, they are relevant for art didactics (Hoelscher, this volume), aiming at new ways of perceiving pictures, interacting with pictures and thinking about pictures.

References

ARNHEIM, R.: *Visual thinking*. Berkeley/Los Angeles [University of California Press] 1969

- BALLARD, H. D.; M. HAYHOE.; P. K. POOK; R. P. RAO: Deictic Codes for the Embodiment of Cognition. *Behavioral and Brain Sciences*, 20, 1997, pp. 723-767
- BARSALOU, L. W.: Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 1999, pp. 577-660
- BRANDT, S. A.; L. W. STARK: Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9, 1997, pp. 27-38
- BUSWELL, G. T.: *How people look at pictures. A study of the psychology of perception in art*. Chicago [University of Chicago Press] 1935
- CHAFE, W. L.: The deployment of consciousness in the production of a narrative. In: CHAFE, W. L. (ed.): *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, NJ [Ablex] 1980, pp. 9-50
- CHAFE, W. L.: *Discourse, Consciousness, and Time. The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago/London [University of Chicago Press] 1994
- ERICSSON, K. A.; H. A. SIMON: Verbal Reports as Data. *Psychological Review*, 87, 1980, pp. 215-251
- FINDLAY J. M.; R. WALKER: A model of saccadic eye movement generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences*, 22, 1999, pp. 661- 674
- FINKE, R. A.: *Principles of Mental Imagery*. Cambridge, MA [MIT] 1989
- GARETT, M.: Levels of processing in sentence production. In: BUTTERWORTH, B. (ed.): *Language production*. London [Academic] 1980, pp. 177-220
- HENDERSON, J. M.; F. FERREIRA (eds.): *The integration of language, vision, and action: Eye movements and the visual world*. New York [Psychology Press] 2004
- HESSLOW, G.: Conscious Thought as Simulation of Behaviour and Perception. *Trends in Cognitive Science*, 6, 2002, pp. 242-247
- HOLSANOVA, J.: Picture Viewing and Picture Description: Two Windows on the Mind. Doctoral dissertation. *Lund University Cognitive Studies*, 83, 2001
- HOLSANOVA, J.: *Discourse, vision and cognition*. Amsterdam/Philadelphia [Benjamins] 2008
- HOLSANOVA, J.; B. HEDBERG; N. NILSSON: Visual and Verbal Focus Patterns when Describing Pictures. In: BECKER, W.; H. DEUBEL; T. MERGNER (eds.): *Current*

- Oculomotor Research: Physiological and Psychological Aspects*. New York [Kluwer/Plenum] 1998
- JOHANSSON, R.; J. HOLSANOVA; K. HOLMQVIST: What Do Eye Movements Reveal About Mental Imagery? Evidence from Visual and Verbal Elicitations. In: BARA, B. G.; L. BARSALOU; M. BUCCIARELLI (eds.): *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Mahwah, NJ [Erlbaum] 2005, pp. 1054-1059
- JOHANSSON, R.; J. HOLSANOVA; K. HOLMQVIST: Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. In: *Cognitive Science*, 30 (6), 2006, S. 1053-1079
- JOHANSSON, R.; J. HOLSANOVA; K. HOLMQVIST: (forthc.): Using Eye Tracking Methodology as a Window to Inner Space. In: PARADIS, C.; HUDSON, J.; MAGNUSSON, U. (eds.): *Conceptual Spaces and the Construal of Spatial Meaning. Empirical evidence from human communication*. Oxford [Oxford University Press]
- JUOLA J. F.; D. G. BOUWHUIS; E. E. COOPER; C. B. WARNER: Control of Attention around the Fovea. *Journal of Experimental Psychology – Human Perception and Performance*, 17 (1), 1991, pp. 125-141
- JUST, M. A.; P. A. CARPENTER: A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 1980, pp. 329-354
- JUST, M. A.; P. A. CARPENTER: Eye fixations and cognitive processes. *Cognitive Psychology*, 8, 1976, pp. 441-480
- KOSSLYN, S. M.: *Image and brain*. Cambridge, MA [MIT] 1994
- LAENG, B.; D.-S. TEODORESCU: Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science*, 26, 2002, pp. 207-231
- LINDE, C.; W. LABOV: Spatial network as a site for the study of language and thought. *Language*, 51, 1975, pp. 924-939
- MAST, F. W.; S. M. KOSSLYN: Eye movements during visual mental imagery. *TRENDS in Cognitive Science*, 6 (7), 2002, pp. 271-272
- NORDQVIST, S.: *Kackel i trädgårdslandet*. Bromma [Opal] 1990
- NOTON, D.; L. W. STARK: Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision research*, 11, 1971a, pp. 929-942
- NOTON, D.; L. W. STARK: Scanpaths in eye movements during perception. *Science*, 171, 1971b, pp. 308-311

- OLSHAUSEN, B. A.; C. KOCH: Selective Visual Attention. ARBIB, M. A. (ed.): *The handbook of brain theory and neural networks*. Cambridge, MA [MIT] 1995, pp. 837-840
- POSNER, M. I.: Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32, 1980, pp. 3-25
- PYLYSHYN, Z. W.: Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80 (1/2), 2001, pp. 127-158
- SNEIDER, W. X.; H. DEUBEL: Visual attention and saccadic eye movements: Evidence for obligatory and selective spatial coupling. In: FINDLAY et al. (Hrsg.): *Eye movement research*. Amsterdam et al. [Elsevier Science B.V.] 1995
- SOLSO, R. L.: *Cognition and Visual Arts. A Bradford Book*. Cambridge, MA/London [MIT] 1994
- SPIVEY M. J.; M. TYLER; D. C. RICHARDSON; E. YOUNG: Eye movements during comprehension of spoken scene descriptions. In: GLEITMANN L. R.; A. K. JOSHI (eds.): *Proceedings of the Twenty-second Annual Meeting of the Cognitive Science Society*. Mahwah, NJ [Erlbaum] 2000, pp. 487-492
- SPIVEY, M.; J. GENG: Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychological Research*, 65, 2001, pp. 235-241
- THEEUWES, J.: Visual selective attention: a theoretical analysis. *Acta psychologica*, 83, 1993, pp. 93-154
- THOMAS, N. J. T.: Are Theories of Imagery Theories of Imagination? An Active Perception Approach to Conscious Mental Content. *Cognitive Science*, 23 (2), 1999, pp. 207-245
- UNDERWOOD, G.; J. EVERATT: The Role of Eye Movements in Reading: some limitations of the eye-mind assumption. In: CHEKALUK, E.; K. R. LLEWELLYN (eds.): *The Role of Eye Movements in Perceptual Processes*. Advances in Psychology, vol. 88. Amsterdam [Elsevier Science B. V.] 1992, pp. 111-169
- YARBUS, A. L.: *Eye Movements and Vision* (1st Russian edition 1965). New York [Plenum] 1967