

Understanding text and picture content as a unity

By Niels Heie

Graphic Designer, researcher, The Graphic Arts Institute of Denmark

In graphic communication messages are often understood in another way by receivers, than intended by senders. Inspired by the idea, that rhetorics is a study of misunderstanding (Richards, 1939), this paper present an experiment in which the combination of pictures and texts as used in modern graphic products is examined. 24 adult informants should choose between sentences related to pictures the way the two are related, for instance as photos and captions in newspapers. 12 pictures, each one with three sentences to choose among, were presented in two different layouts: L1 with one picture, L2 with three copies of the same picture. Informants were eyetracked while deciding which combination contentwise could be considered most objective, subjective, non-sensical.

Time measures showed that informants used app. 25% more time on the subjective than on the objective task.

A difference in the subjective task was found between the two layouts: in spite of the average time used was the same, in L2 layouts informants used more time at pictures than at texts, opposite the result when choosing the subjective combination in L1 where most time was used on text.

App. 92% of choices for the objective content relation were as foreseen, app. 77% for the subjective.

The experiment showed that interdependencedness should be considered when making visual communication and that both layout and complexity of content influences the time used by the reader. For further research it is suggested that pictures and texts in graphic communication products are seen as elements in the argumentative structure introduced by Stephen Toulmin.¹

Two codes

Multimodality is at the heart of modern graphic communication products. People working with graphic communication, being author or designer, photographer or advertiser works with multimodality, more or less consciously.

In this article multimodality stands for the use of static pictures and texts presented in static layouts. The contents of the elements are considered interrelated as a way to present a message.

Multimodality is used in wide variety in graphic products, from simple compositions as in identitycards containing a photographic portrait with a name under it, over more complex compositions of picture, title, authorname and tagline on a bookcover, to broadsheet newspaper spreads containing several articles with several pictures combined by layout with headlines, captions, copytexts.

Multimodality is used whether graphic products are intended for efferent reading as in an IKEA do-it-yourself manual or for aesthetic reading as in a lifestyle magazine (Rosenblatt, 1978).

Pictures and texts are signs. In the semiotic terminology of Charles S. Peirce (Lizla, 2001) texts are symbolic signs, pictures are iconic.

The combination of pictures and text as a means of visual expression is in line with Paivios concept of dual coding, dividing in cognitive processes in two systems. *The imagery system* is understanding scenes and situations and generates images. The verbal system is understanding language and generates speech (Paivio, 1978).

Elements in visible messages

Communication assumes that two parties, sender and receiver, agrees on the meaning of the signsystem used. To be communicated a message has to be encoded in a way that the sender imagines the receiver is able to decode.

Graphic communication using multimodality is dependent on the readers ability to read written language, but also his ability to read pictures, as well as to read the composition of the two shaped by the layout.

In spite of this, not much research have been done in the field (Rayner 2001, Holsanova 2005), all though, for instance, lots of children have learned to read text looking at words placed in a multimodal compositions: A word is related to a picture to make the child understand, that the referent of the word is presented in a picture of the referent. Children learn that bird is the word for a picture of a bird – which makes Rene Magrittes painting *"The Use of Words"* (1928), of an iconic pipe above a symbolic line saying *"This is not a pipe"* a basic picture of a basic condition.

Text is written words that the reader understands as pictures of spoken words. Text has a message. Where as pictures – in this study photographic pictures – can be considered as a medium without the linguistic quality of having a message (Barthes, 1977, Elbæk, 1987). Multimodal graphic communication products then, combine language and non-language elements in compositions as if they could be understood as unities. But can they?

Reading both picture and words

Elbro (Elbro, 2001) in a model of reading ability defines it, as the ability to decode letters and words related to the ability to understand the meaning of language.

To fit the multimodal communicating products then, the concept of reading must be expanded because a reader has to decode both text, pictures and the layout containing them and, through the cohesiveness of the elements, understand the meaning of elements and layout as a coherent unity: i.e. why *this text* goes with *this picture* within *this format* in *this way*.

The expanded model thus looks like this: reading ability is the decoding of visual elements and the understanding of there mutual relation as a whole. The model suggests that a dual coded message should be read in a dual coded way and result in a coherent understanding of the meaning of the presented modalities.

If a message is understood in a certain way when presented in text alone, what happens to the understanding of the message if it is presented in the same text but added a sense making picture? Taking the multimodal version as a coherent unity, the difference in understanding between the unimodal and the multimodal version ought to be a difference in the content of the readers inferring.

Background for better routines

In the production of graphic communication experience based habits dictate daily routines on how to combine pictures and words. To get more explicit knowledge in the field, i.e. get a better understanding of picture/text cohesion and eye movements, we made a study using eyetracking equipment to test whether combination of pictures and words in this daily-basis way, were understood by informants as intended.

Our experiment was in line with two specific studies, one about the reading of illustrated textbooks, one about the understanding of ads composed of pictures and texts having either implicit or explicit content.

In the illustrated textbook study (Hannus, 1999) six spreads from new biology textbooks were presented to 108 fourth grade schoolchildren scored to be on different levels of text reading ability. The experiment showed, that students in general mostly learned from text, that students of high text reading ability increased their understanding of the content by use of the illustrations and that students of low text reading ability were confused by the illustrations. Eyetracking showed that able readers could change from text reading to picture reading and back again without confusion. Less able readers had problems changing from one modality to the other.

In the ad study (Radach et. al. 2003) 32 grown ups were presented to several ads composed of pictures and texts. The contents of the ads were defined as explicit or implicit: in

explicit ads content of text and illustration was in a direct semantical relationship and the message was presented in a straightforward way. In implicit ads the content was in a more complex semantical relationship and the message was presented indirectly. The result was that explicit ads were read faster than implicit ones and that informants expressed a bigger "liking" for implicit ads than for explicit ones. The *likeability* part of the result is interesting. *Likeability* in ads correlates to the sale of advertised goods (Radach et. al. 2003).

In a not published small pretest-experiment we tested if illustration had any influence on solving a task, where three 4-person groups of grown ups of expected equal reading abilities (attending same level of education) individually should decide, which word six letters non-obvious presented were forming. The letters were presented on a computer-screen this way:

A
V U
A L T

The word to find was *valuta*, i.e. danish for *currency*. To get an impression of the task read the word currency in this layout:

Y
C R N
U R E C

One group saw letters alone, another group saw letters in a layout with a detached photo of coins, a third group saw letters with a detached photo of keys. (Detachment of a photo removes the background and makes the appearance of the figure letterlike). While looking informants eye movements were registered.

Result, average time used to find the answer:

- coin-photo-group 09 sec
- no-photo-group 42 sec
- key-photo-group 52 sec

The conclusion of these studies and the pretest was:

- Readers overall understanding is influenced by the picture-text content relationship
- The picture-text content relationship can be described as explicit or implicit
- Time used seems to vary according to difficulty of task

The not so simple experiment

Overall we wanted to test how informants would choose among a series of picture and text combinations presented four times and if choices were as researchers expected. Informants should choose as if they were like an editor choosing adequate pictures and captions.

Three choices should be made within the frame of the explicit/implicit concept. To get a better understanding of the limit of the explicit/implicit categories, we invented a category, call it "nonplicit", i.e. if the picture/texts relation was experienced as absurd.

Before the test, informants saw the two layout versions of the stimulus pictures, were informed of the task and were interviewed on their familiarity with the two concepts *objectivity* and *subjectivity*. These words were more common and were used as synonyms for explicit and implicit. Articles in The Danish Encyclopedia were background for a mutual understanding of the two concepts. For non-plicit we used non-sensical.

Further they were told to use as much time as needed and to say out loud the number of the text chosen for the combination, before tapping space on a computer keyboard placed in front of them to call the next stimulus picture to the screen.

The 24 randomly chosen adult informants were divided in 4 groups A, B, C, D. Group A + B saw a set of text elements related to picture 1-6, but A was presented to a layout different from B. Group C + D saw a set of text elements related to picture 7-12, also in two different layouts.

The 12 photos were randomly chosen among newer pictures in Danish newspaper Politikens archive taken by its photographers and were not known by informants.

The motives varied, from women giving each other massage in a public bath to a row of windows in an empty sunlit living room.


Each photograph was presented with one out of four questions and three lines that would go with the photograph, as in a headline or a caption. These 36 texts were written by the researchers based on their common understanding of picture/text combinations.

The questions were if the combinations of picture and text could be considered respectively objective, subjective, non-sensical and emotional arousing.

In sum: Each serie contained 24 stimulus pictures. Each stimuluspicture contained one out of six photos, one out of four questions, three out of 18 texts to choose among.

Example L1 (group A):

Hvilken tekst er mest OBJEKTIV?



- 1) Et lille hus på landet
- 2) Fængsel for livstid
- 3) 32 tomme bakker

Top left **A question**, here: *Which text is the most objective*

In the middle **A photograph**, here: *A little house in the landscape*

Under it **three lines**, here: *A little house in the landscape, Prison for life, 32 empty trays*

Example L2 (group B):

Hvilken tekst er mest OBJEKTIV?



1) Et lille hus på landet



2) Fængsel for livstid



3) 32 tomme bakker

Same content, but photos shown three times, with the captions beneath them.

If group A saw the objective question in layout L1, the subjective question was shown in layout L2. If group B saw the objective question in layout L2, the subjective question was shown in layout L1. Etc.

This rather complicated structure of stimulus pictures was decided upon because pre-test showed, that if all stimulus pictures had the same layout the experiment would result in eye movement patterns that were very alike and would be difficult to analyze. Recognition would be too important.

While looking at the pictures informants eye movements were eyetracked with SMI equipment using iView 3.01 software. Infrared light was the source to create eyemovement reflections that were registered by a Red II SMI camera, 50 hz. Software Image Stimuli 1.0 was used for randomizing the order of stimulus pictures and to discriminate data.

Results, some of them surprising

16 out of 24 datafiles each containing 24 datasets were considered valid, having eyetrack data for 90% of the time spent on looking. Data from the experiment were analyzed by dividing each stimulus picture in *Areas of Interest* (AOI): the question was one AOI, the picture another, each of the caption possibilities one AOI each. Stimulus pictures with one photo contained five AOIs, stimulus pictures with 3 copies of the photo contained seven AOIs.

Then we found out how much time informants used on each AOI and related it to the time used on the stimulus picture as a whole. 2304 AOI's were scored, 144 per stimulus picture. Most of the results were t-tested.

First hypothesis suggested that there should be no empty AOIs if both photos and words were read in order to answer the question. It was not found.

118 (5%) of AOI's were empty. The reason can be that the photographic motives and the captions possibilities are seen four times, why recognition will result in an extended use of the parafoveal sight. Another factor can be the difference in layout. Where captions are placed close to the photos (L2) text might be read through parafoveal sight (Rayner et. al. 2003).

Our hypothesis on difference between data in the objective and subjective questions was in line with (Radach et. el, 2003) that informants would spend more time on the more difficult subjective task than on the easier objective task. Further that time spend on looking at the photos would show parallel differences.

Average amount of time used:

L1	Objective questions	2.4 sec
	Subjective questions	3.0 sec
L2	Objective questions	2.2 sec
	Subjective questions	3.1 sec
L1+L2	Objective questions	2.3 sec
	Subjective questions	3.05 sec (t=0,002882)

Differences average amount of time used on photos:

L1	Objective questions	0.8 sec
	Subjective questions	0.9 sec (t = 0,1678)

Task related differences in average amount of time used on photos:

L2	Objective questions	0.9 sec
	Subjective questions	1.4 sec (t = 0,0092)

Differences in average amount of time used looking at photographs in the two layouts was only significant for the subjective question in L2. Here informants used more time looking at photos. Interesting enough the overall time used was not influenced by this, which means that informants have used less time on the words.

On predictability the analyze was done like this: finishing the production of the 24 stimulus pictures, researchers registered there predictions of informants solutions. All though the procedure was satisfactory pretestet – result was in line with hypothesis – it cannot be said to be valid. Even so, we repeated it in the experiment, because it could give us information to use in further research.

The result of comparing the list of predictions with informants solutions showed that researchers and informants agreed on:

L1	Objective questions	98%
L2	Objective questions	92%
L1	Subjective questions	80%
L2	Subjective questions	75%

The prediction system was also used on the question of nonsensical relations between pictures and texts. Here the result was that researchers and informants agreed on:

L1	Nonsensical questions	87%
L2	Nonsensical questions	83%

The results can be concluded like this:

It makes sense to work with the categories objective, subjective and presumably nonsensical combinations of photos and words in the form used here.

Informants spend app. 25% more time on the task to decide which of the possible texts are connectable to the photo when their interconnectedness (Richard, 1939) could be described as subjective, compared to objective. The result is parallel to (Radach, 2003) in that the amount of time used on reading implicit ads is bigger than on explicit ads. Amount of time used to decide the subjective picture-text relation is bigger than deciding the objective picture-text relation.

(Here it should be noticed that the time used on answering which combinations were nonsensical were app. as the objective task).

The difference in layouts resulted in different eye movements strategies only when informants had to decide the subjective combination and the layout contained three identical versions of the photo (L2).

Although total time used on the subjective question was the same in both layouts, informants used 50% more time looking at photos in L2 than in L1:

L1	Total 3.0 sec / Photo 0.9 sec
L2	Total 3.1 sec / Photos 1.4 sec

Difference in time when informants looked at photos in both L2 tasks was:

L2	Objective questions	0.9 sec
L2	Subjective questions	1.4 sec

($t=0,0079$)

When the subjective question in L2 results in such a different eye movement strategy compared to the objective L2 question and the two L1 versions of the task, it may have to do with the character of the task. Subjectivity is by definition different from objectivity: it is not as predictable. That task triggers what to look for, was found by Yarbus (Yarbus, 1967). So being shown three copies of the same photo with three different sentences and the task is to decide which combination is the subjective one, can in itself easily be understood as a task constructed in a subjective way. And if the photos in fact *are* the same, can only be decided upon by informants after severe investigation. Because, why show a layout containing three pictures that are alike, when one is enough?

Names, a bridge between pictures and texts

To extend knowledge that can be used on a daily basis by people working with graphic communication more detailed experiments should be done. But a better understanding of the interconnectedness of the two modalities on a subordinate level can be advantageous.

Here are some arguments for an approach, that can be tested in further experiments:

As said earlier picture and words combined is a match of two elements that are different in that one is language, the other is not. To find what unites them one have to find what they have in common.

Pictures like photographs are visual structures containing one or more perceptual figure/ground relation (Köhler, 1947) that by likeness, i.e. iconicity, can be understood as

signs for visible real world elements spacial related. Visible real world elements, things, often have names.

Names are words, sentences are composition of words. So when a picture contains elements that can be named, a sentence presumably can be constructed by the names of the figure and the names for their relation to each other. Thus a picture showing a farmhouse in a landscape, can be described in the sentence *a picture showing a farmhouse in a landscape*.

According to Frege a sentence has meaning and reference (Frege, 1993). If it has meaning it is logically cohesive and proper constructed according to syntactic rules. But it do not need to have reference, i.e. there do not have to be a referent to the sentence. If there is a referent to the sentence *a farmhouse in a landscape* it is because there actually is *a farmhouse in a landscape*. To Frege the reference/referent relation is the thruth condition of the sentence.

So, what is the difference between the content of the sentence containing the words *A farmhouse in a landscape* and the content of a photo containing figure/ground gestalts that can be named with the words *A farmhouse in a landscape*? The difference is that the sentence claims a class, where as the photo shows an entity. If the picture describing sentence and the caption sentence is combined in one, the result is a repetition of the sentence, a tautology: *A farmhouse in a landscape, a farmhouse in a landscape*. The sentence make sense if its is understood as *A farmhouse in a landscape (is) a farmhouse in a landscape*.

By abduction the combination of picture and text can be understood to contain the meassage: the house of the sentence is the house of the photo. In that way a reader understands the content of picture and words as a unity. The message of a picture/words unity then is, that the sentence is mentioning one out of X members in a class and that the mentioned one is the one shown in the picture.

The content of this combination can be defined as objective.

If the composition consists of the picture of the farmhouse combined with the sentence *Prison for life* the picture sentence and the caption sentence is saying: *A farmhouse in a landscape, prison for life*, which makes sense if one abducts the meaning to be: *A farmhouse in a landscape (can be) prison for life*, because a farmhouse may be a prison to be in and thus an experience. Experience is a subjective quality. To understand other peoples experience empathy is needed. As shown in our experiment, it takes time to understand subjectivity.

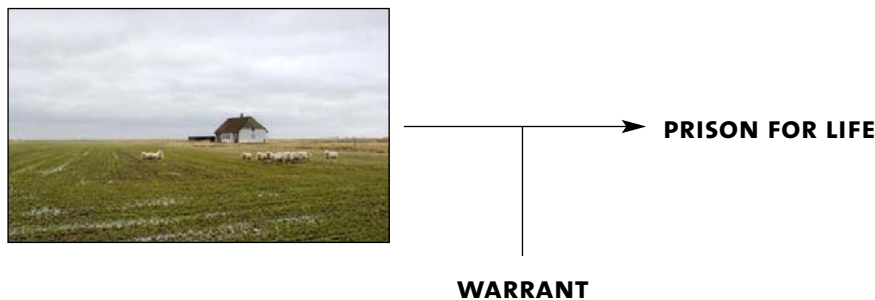
If the combination consists of the picture of a farmhouse and the sentence *32 empty trays*, the two make the sentence: *A farmhouse landscape, 32 empty trays*. It is difficult to abduct a unifying meaning of the two. It do not make sense to understand the combination of sentences either as: *A farmhouse in a landscape (is) 32 emty trays or (can be) 32 empty trays*.

Logic, pictures and text

The content of a message is a claim. According to Stephen Toulmin (Toulmin, 1958) a claim is one of three vital elements of an argument. The other two are: data and warrants. The claim is what the argument contains of new information, the data is the basic upon which the claim is made and the warrant is the background for the claim related to the data. The triadic structure of the argument is like this: *(data) so (claim) since (warrant)*:



The structure can be transformed into the structure for a message: *(data) is (claim) since (warrant)*. If the argument structure is used on a combination of picture and text like the ones in the experiment described here, the result is like this:



The picture-text combinations used can be understood as elements in arguments only, if one understands that there is no warrant. But in the objective and the subjective task the data content of the picture and the claim of the words have a character that makes it possible for the reader to infer, what the content of the warrant could be if data, claim and warrant should form a proper argument, namely:

- The objective combination:
What you see is a picture of what is written.
Data in the picture make it probable that the sentence has a referent, warrant content must be about the house.
- The subjective combination:
Living in a house can be experienced as being in prison.
Data in the picture make it probable that the sentence has a referent if a house like the house in the picture can be experienced as a prison. Warrant content must be about being in the house.
- The non-sensical combination:
32 trays can not be united in an argument with data containing the appearance of a house by a warrant, unless one infer the warrant to be the denial: there is nothing in favour of claiming that a farmhouse in a landscape should do as data for a claim of 32 empty trays.

Most graphic products are read like that. People infer from headlines and pictures what the warrant content in the form of for instance copytext or logos can be.

To interpret like this the way readers understand the content relation between pictures and words, can expand the understanding of the difference in eye movement patterns that Yarbus (Yarbus, 1967) found in the experiment, where respondents should look at a painting of a scene with people in a room and answer different questions related to the picture. The difference in eye movement patterns can be described as the search for data that can relate to a claim. For instance can the question: how old are the persons, be understood as the claim: age is visible in peoples appearance, what age can one infer from the data in the appearance of the persons in the picture? Although Yarbus' informants were asked in spoken language, whereas informants in our experiment were asked in written language, the logical content of the experiments is the same.

This basic understanding of the unity of picture and words as elements in an argument has to be tested in new experiments using eye tracking. Hypothesis is, that a logical

approach to multimodality like this can lead to a better understanding of readers eye movement patterns and expand the understanding of cohesiveness and of integrative saccades (Holsanova, 2007). Understanding messages as arguments has the advantage of being both a tool for research analyzing products and experiments on a subordinate level and a tool in the production of graphic communication products, for instance when sketching new products or testing prototypes.

¹ Part of the empirical data from the experiment is not used in this paper. The experiment was made in 2005 by *Niels Heie, Karen Margrethe Oesterlin* and *Frank Christensen* from The Graphic Arts Institute of Denmark. It was supervised by prof. *Kenneth Holmqvist*, University of Lund, Sweden.

- Barthes, Roland*: Image, Music; Text. New York, 1977
Elbro, Carsten: Læsning og læseundervisning, København 2003
Elbæk, Ib: Er billeder sprog? Rolig papir 59/97, Roskilde 1998
Frege, Gottlob: On Sense and Reference, in Meaning and Reference, Oxford 1993
Holsanova, Jana et. al.: Tracking information of text & pictures, Lund 2007
Hannus, Mattis, J. Hoyna: Utilization of Illustrations during Learning of Science Textbook Passages among Low- and High-Ability Children, in Contemporary Educational Psychology 24, (1999)
Köhler, W: Gestalt Psychology, New York 1947
Lisza, James Jakób: A General Introduction to the Semeiotic of Charles Saunders Peirce, Indianapolis 1996
Paivio, Allan: A Dual Coding Approach to Perception and Cognition, Lawrence Erlbaum Associates 1978
Radach, Ralph: Eye Movements in The Processing of Print Advertisements, Elsevier Science BV 2003
Rayner, Keith et al.: Integrating Text and Pictorial Information: Eye Movements When Looking at Print Advertisements, Journal of Experimental Psychology 2001
Richards, I. A.: The Philosophy of Rhetorics, London 1936
Rosenblatt, Louise: The Reader, the Text, the Poem, Illinois 1978
Toulmin, Stephen: The Uses of Argument, Cambridge 1958
Yarbus, A. L.: Eye Movements and Vision, New York 196