Integrated processing in multimodal argumentation

Jiang Wangqi, Peking University
Paul van den Hoven, Tilburg University, Utrecht University

Introduction

The question addressed in this paper is simple. If the argumentative function of a multimodal narrative text requires the integration of the information from different modes, among which verbal ones, what model for the order of processing and the integration of information do we need to adopt? Using discourse analysis as a method, we will argue that such a model needs to meet a number of requirements.

We will argue that at least partially we are still in the realm of linguistic connectionism. By that we mean that the processing of the verbal modes is cognitively connected with and is influenced by the processing of other modes, perhaps similar to what has been argued that is the case in the multimodal processing of speech and gesture. The integration of for example a verbal speech voice-over mode with a pictorial mode is not a post-perception, late, dominantly conscious inference process that operates on more or less finalized mental representations of the separate modes. It depends on perceptual paring, on stipulated cross-modal congruence, and on cross-modal interactions between initial perceptions and knowledge based inferences. We will show how a unitary source assumption leads to cross-modal temporal paring as well as to the projection of the structure of one mode on the other mode. A model needs to allow such cross-modal connections to account for the fact that for example simultaneously appearing forms in two modes are interpreted as a cross-modal metaphor in which the structure receiving mode is taken to be the source domain.

We will also argue that the integration of preconscious bottom-up information and inferences from top-down information across modes is characterized by a complex blending of structures. Observations from two brief fragments of multimodal narrative argumentative texts are explained by postulating the possibility of cross modal integrations in which either (1) elements from the one mode fill slots in the dominant structure of the other mode, or (2) integration takes place because the structures of the information from both modes is (taken to be) congruent, or (3) the structure of the information of the one mode is ‘used’ to structure the other mode, or (4) structural elements from both modes are blended into a new structure.

More restricted models - we claim - cannot explain these relatively straightforward examples of multimodal discourse processing. This underpins the validity of the theoretical insight that (multimodal) discourse processing is a holistic process in which the meaning attached to the text is more than can be computed out of its locally restricted structures.

Holism in discourse processing

In Eastern as well as in Western discourse theory a tradition can be discovered of a holistic approach to the processing of discourse (Wangqi 2011). Especially in the Chinese theoretical tradition, such an approach has been dominant. It should be called *holistic* because it recognizes that the meaning extracted from a text is always more than what can be computed from the conventional meanings of its parts and their internal structural relations. At the very beginning of the processing of a text, often even before the actual start of the unfolding of the textual sign vehicle, thematic and formal structures, scripts, scenario's are activated that guide and influence the process of interpretation. Besides the widely recognized top-down element in discourse processing, it is also the integration of different inputs that makes the processing a holistic process. Peircean *semiotics* (compare Van den Hoven 2010a, 2010b), Longacre's *Text linguistics* and the 19th century work of Liú Xīzāi, *A Survey of Essays on the Classics* and in recent times *conceptual integration theory* (Fauconnier & Turner 2002) all try to explain how the integration of such diverse top down and bottom up inputs in an advancing incremental process may lead to a meaning attached to the textual sign vehicle that is fully intelligent and explainable afterwards, but that cannot be computed beforehand from a limited set of general rules for the computation of meaning.

With regard to the Chinese linguistic tradition, there are two important theories: the formal, structural theory of qǐ 起, chéng 承, Zhǔan 转 and hé 合 (initiation, elaboration, transition and conclusion), and the semantic, content theory of zhǔtí lùn 主题论 (theory of thesis). This tradition, dating back to 501 when Liú Xié 刘勰 published his Wénxīn Dīaolóng 文心雕龙, was finalized by Liú Xīzāi 刘熙载 in the 19th century. On the basis of the developments by many scholars over the years, Liú Xīzāi argued in his Jīngyì Gài, 经义概, *A Survey of Essays on the Classics*, that "In any essay, the thesis must be capable of being summarized in a word. The thing that can be expanded into hundreds and thousands of words on the one hand and contracted into one word on the other is no other than the thesis.[…]. The thesis must be expandable and contractible at the same time. It must be related to the main ideas of each chapter, section and sentence, the former cannot be separated from the latter." (Liú

1991, 14). He explicitly combined the discussion of qǐ, chéng, Zhǔan and hé with that of the thesis. In his view "In the sections of Pòtí 破题 and Qǐjiǎng 起讲, one must grasp the thesis firmly. Then in the sections of Chéngtí 承题 and Bābǐ 八比 one can elaborate on the thesis in detail" (Liú 1991). He emphasized the importance of respecting the topic. "One should start to respect the topic right from Pòtí and Qǐjiǎng […]. To respect the topic means to make it clear how important the topic is, and as a result the whole essay turns out to be a worthy essay" (Liú 1991, 21). He also talked about the specific ways to explain the title, suggesting that Pòtí itself should be seen as a whole essay and consists of the four important components of qǐ, chéng, Zhǔan and hé as well. As we will see, these ancient theories projected on multimodal discourse, precede the unitary source assumption on the post-perception level as well the background for the obvious assumption that interpreters make that cross-structural projections are to be expected form the assumption that coherent argumentative discourse develops one thesis in an intelligible sequence (though a sequence can also be a multimodal simultaneous presentation).

Argumentative narratives as a discourse type can be expected to manifest these holistic characteristics. The narrative structures as well as the argumentative structures add to the thematic structures that guide the interpretation process. And, as dialectic argument theories (Van Eemeren & Grootendorst 2004) demonstrate, also the wider context of the often ongoing discussion will frame the processing of the text. This makes multimodal argumentative narrative discourse an interesting object from a theoretical cognitive linguistics point of view. In this paper we will present two short fragments of this specific kind of multimodal discourse to argue the holistic claim. Our findings are that structural characteristics from the one mode seem to transfer to the other mode in a flexible way and that indeed there are indications that preconscious perceptual mechanisms are connected with post-perceptual inference.

Fragment 1

The first example comes from a video clip, produced by George Henry Aulson IV as his victim impact statement, presented during the probation hearing of the convicted murderer of his father (http://www.youtube.com/watch?v=uATxCVAAik0). Although certainly not the most spectacular example of multimodal integration, we want to illustrate the major theoretical issues with a 26 seconds fragment from this almost 9 minutes long video. In this fragment a voice-over, apparently George Henry Aulson, speaks the words:

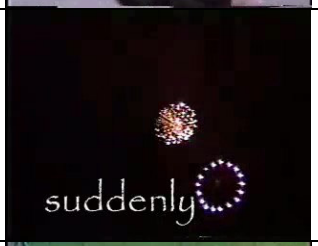
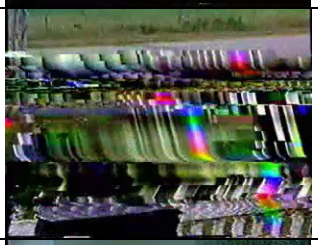
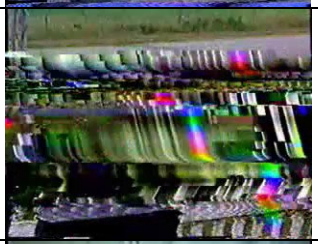
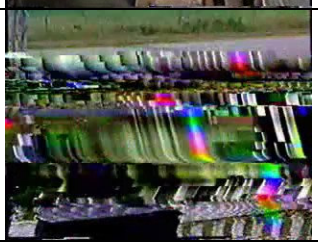
> On July 7th 1991 my father was suddenly taken away from me. I was only 5 years old. He didn't die in his sleep or from cancer but from a horrific and fatal stab wound in the chest, suddenly, damned quickly, without warning and unexpectedly. I wanna emphasize this. Why? Because I never got the chance to say goodbye to him. Would I feel better if I've had a chance to? I don't know but I wish I did.

The transcript of the verbal speech is intelligible as a narrative with a (potentially) argumentative function, without the accompanying music and visual elements. From the verbal speech alone the interpreter can construct a narrative. The initial disturbance that puts the narrative dynamics in motion is the murder on the protagonist's father. The interpreter can understand this narrative as an argument that regards the liability of the murderer for at least taking from the protagonist the opportunity to have a final goodbye with his father.

By contrast the visual mode is not intelligible as coherent discourse on its own, nor is the structure of this mode predictable from the verbally presented narrative. But it is interpretable in its symbolism, also because of a series of projected words. During the almost nine minutes of the video we see: fatherless - suddenly - I don't know - dead - wake - murdered - emotions - my dad should be here - dream - a reassurance - LIFE IS GOOD - never - why. The pictorial sequence is full of fragments of happy moments with father, full of symbols of goodbye such as hand waving, constantly interrupted with a kind of visual noise, and several times we get a fade-out into a lengthy, black moment. The interpretation is guided however by the 'thesis' in the verbal mode. Here we see the (normative) theory of Liú Xīzāi illustrated as the ground for interpretative dynamics. As he declares about the thesis: "It must be related to the main ideas of each chapter, section and sentence". We can add: it seems to be expected - under conditions to be specified - it is expected to be related to each mode.

The other dominant mode, the music, is very significant for the intended (American) audience: *Adagio for Strings, opus 11* of Samuel Barber, which was also played during the memorial for the 9/11 victims on September 15, 2001. The video was produced after that. We will neglect this mode here, but, according to Cohen's congruence associationists framework model (Cohen 2001), the integration of this mode needs to be accounted for in such a way that it may select a focus on structurally congruent elements in the other modes, even without being heard acoustically (= be consciously present in STM).

We present the visual and speech mode (unjustly neglecting the prosodic and musical mode) in a simplified score.

time	visual mode	voice-over
1.26		On July 7th 1991 my father was
1.28		suddenly taken away from me. I was only 5 years old.
1.30		He
1.30		didn't die in his sleep or from cancer
1.33		but
1.33		from a horrific and fatal stab wound in the chest
1.37		

1.37		suddenly
1.37		damned quickly, without warning and
1.41		unexpectedly … I wanna emphasize this.
1.47		Why? Because I never got the chance to say goodbye to him. Would I feel better if I've had a chance to?
1.50		I don't know but I wish I did.

An analysis of this fragment enables us to argue the plausibility that the processing of the perceptual modes is connected on several levels, in a diversity of manners. We will present here a limited set of observations that suffices to argue the complex blending of modal structures, based on preconscious cross-modal integration.

Preconscious cross-modal perceptual paring seems to operate on these artificial multimodal artifacts in a way that is similar to perception in non textual situations. Cross-modal perceptual paring is a documented form of preconscious interpretation (Bertelson & De Gelder, 150). The interpreter of multimodal inputs makes an early assessment of the degree of concordance of the total input; default is that the modes are interpreted as congruent (simultaneous and unilocal, for example) and of a unitary source which leads to a cross-modal pairing of the perceptions. If the assumption of a unitary source is strong, disturbances in the congruency of the sign vehicle result in recalibrations of the perceptual

system. These recalibrations lead temporarily to after-effects. This mechanism has been demonstrated for the cross-modal perception of proprioceptivity and vision, for the cross-modal perception of vision and sound (the so called ventriloquist effect) and for the interpretation of audio-visual speech (the so called McGurk effect).

In our fragment the audio-speech mode (voiceover) is presented together with a visual mode, obviously in an intentionally constructed textual artifact. The ventriloquist effect still seems to occur. Even though there is no talking head visible, the interpreter perceives the audio source as somewhere in the visual field. However, only in a very specific configuration of the technical equipment, the physical source of the sound of the voiceover will indeed originate from somewhere behind the projected picture. This means that in the interpreter's cognition the localization of the speech source has been shifted into the direction of the visual source.

This ventriloquist effect indicates that a unitary source is strongly assumed. Psychological literature makes a distinction between the effect that originates from automatic, mandatory perceptual processes or from post-perceptual judgmental processes (Bertelson & De Gelder, 151). Evidently we cannot decide in which realm we are here. An empirical criterion seems to be whether recalibrations and therefore after-effects occur. The distinction as such originates from a rather strong assumption of modularity. However, a model that allows for connections between 'automatic' mandatory processes and inferences is plausible. Empirical data that make a strong claim for the pairing concept in preconscious perception are fully compatible with the possibility that learned knowledge about the textual artifact influences the assessment of the degree of concordance, and such data do not forbid that paring phenomena influence inferences in post-perception processing. The observations that we will discuss concern processes that build on this unitary source assumption. That supports our idea that preconscious and conscious processes are connected.

We will show how the unitary source assumption renders disruptions of temporal as well as thematic concordance into input information that activates complex top-down structures. These structures guide the integration of upcoming text elements.

The first example concerns a disturbance of the temporal congruence of the speech structure and the visual structure. In the speech mode, this structure is the prosodic and thematic segmentation. In the visual mode, this structure is the repetition of specific pictorial elements (most important the colorful visual noise in 'shot' 3-5-7 and the black fade-out) and

the segmentation by means of hard-cuts. The unitary source assumption, with the expectation of simultaneous, congruent modes, renders every disturbance meaningful.

The temporal disturbance is at the beginning of the fragment. Just before the fragment starts, a new section is announced in the speech mode: "I am going to present a few stories about how I felt and how it impacted my life". This is followed by a long pause. And then the fragment starts with: "On July 7th [and so on]". This new section in the speech mode however does not coincide with a cut in the visual mode, but it starts a little bit before that cut. This temporal incongruence, which is repeated several times in the movie in a consistent way, seems to support the activation of knowledge about specific discourse relations between the modes. The early start of the speech mode indicates that the voiceover knows the pictorial elements in advance and that in a way s/he controls them.

It is plausible that the temporal incongruence significantly contributes to the activation of this discourse scheme. When we virtually change the incongruence, this leads to a different discourse scheme. When the modes are temporally congruent, the preferential discourse scheme is that of the voiceover as a narrator, perfectly following the coordinating structure of the visual mode. We often see this option in movies for young children and in the expositions in fiction movies. When the visual mode precedes the speech, a discourse scheme is activated in which the voiceover is a commentator who has no or limited knowledge of what is coming next in the pictorial mode; the pictorial mode becomes an interpretation object for the voiceover as well as for the interpreter. This is the case in non-fiction live registrations, such as sports events.

In this video fragment, the visual mode by itself lacks a strong, narrative structure. The discourse situation with its omniscient voiceover however guides the interpreter into an integration process in which the coordinating structure of the speech mode is projected on the visual mode (be it with incidentally the temporal delay in the visual mode to confirm the discourse scheme).

This projection of the speech mode structure on the visual mode makes the moment of appearing of a picture significant, but primarily as a speech increment. Such an integration process is what Fauconnier and Turner describe as the dynamics of a single-scope network (2003, 127f). The structure of one input space (the speech mode) is projected on the blend in which information from several input spaces (speech and visual) is integrated.

So now we see how:

1. perceptual pairing renders cross-modal temporal incongruence significant;

2. cross-modal temporal incongruence activates specific knowledge about the discourse situation;
3. specific knowledge about the discourse situation facilitates the projection of the speech structure on the visual mode;
4. projection of the speech structure on the visual mode stimulates the interpreter to integrate the pictorial characteristics within the speech mode.

On the thematic level point 4 can be illustrated with the fireworks that coincide with the speech mode element "suddenly". It is highly plausible that because of the single scope projection from speech mode to visual mode, a process starts in which knowledge activated by a pictorial element is projected on the (simultaneous) speech. Therefore the SUDDENNESS and the SPEED in the event of the FATHER'S DEATH IS related to FIREWORKS. Fireworks are a domain of happiness – we will go into that soon - but also a domain of BIG BANGS, CHILDREN'S FEARS, A BOLT FROM THE BLUE, THE CLAP IN THE NIGHTLY DARKNESS, and so on. In a word, the dynamics apply that we know from metaphor theory. These include the mutual dynamics between source domain and target domain, framed by the grounding (Lakoff & Johnson 1980, and specifically Coulson for the concept of grounding). The integration according to the steps (1) - (4) makes that the source domain is in the visual mode, the target domain is in the speech mode, while the grounding is the product of the integration processes so far. Fauconnier and Turner rightly analyze metaphor interpretation as the dynamics of a double-scope network (2003, 131f, 280f). So indeed double scope integration works cross-modal and in close connection with the steps (1) - (4). Basically this is what we promised to argue.

This chain of connected cross-modal integration is prolonged even further, now driven by a thematic incongruence. Again such incongruence gets meaning because congruence is presupposed. We did already observe the antithetic element in the fireworks as metaphor for the suddenness of father's death: fireworks are first of all connected with happy festivities. This antithetic relation between the modes is structural. DYING IN THE SLEEP goes with A PEACEFULLY SLEEPING YOUNG CHILD (George Henry?), A HORRIFIC AND FATAL STAB WOUND with A SCHOOL CLASS OF HAPPY CHILDREN WHO RAISE THEIR FINGER, DAMNED QUICKLY with A QUICK WITHDRAWAL OF THESE FINGERS, SAYING GOODBYE TO A DYING PERSON goes with CHILDREN LAUGHING WHILE WAVING GOODBYE.

An analysis of the video beyond this short fragment shows that these antithetic relations are presented time and again. This cross-modal antithetic relation develops the dominantly verbally presented argument. The argument has evidently two lines: the line of

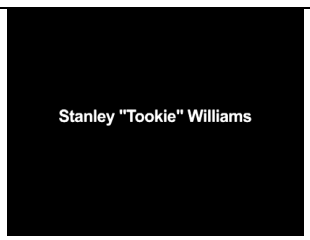
reality: how life is since father's dead, and the hypothetical line: how live would have been if father had not died. The speech mode largely limits itself to the first line. The visual mode largely presents pictorial elements that come from happy moments with father or other happy pictorial scenes. So the hypothetical line is enriched by this cross-modal antithetic relation.

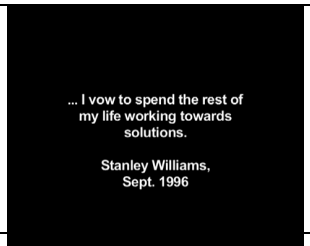
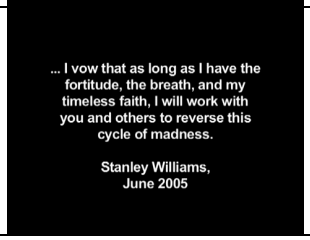
We see a process that switches a back and forth between the modes, developing from preconscious perceptual information about a unitary source up to post-perceptional inferences. In this incremental process we need to assume that information is used immediately after it becomes available. In the process, text structures are integrated in an activated discourse scheme, in a narrative structure and in an argument structure, while world knowledge fills the source domain to 'solve' the metaphorical elements, and a creative projection needs to be assumed from the reality as presented in the speech mode, to the hypothetical world as suggested in the visual mode, back to the integrated cross-modal argument structure.

Fragment 2: Stanley 'Tookie' Williams

The second fragment comes from a petition for clemency for Stanley 'Tookie' Williams (http://www.youtube.com/watch?v=KhFoeJPP6HE) . Stanley 'Tookie' Williams has been convicted to the death penalty. The petition is a last attempt to convince the governor of California not to effectuate the penalty. Stanley Williams is said to have changed his life from a violent gang leader into a Nobel peace prize nominee who has dedicate his life to expelling gang crime. The petition has not been successful.

Here is a simplified score of the first minute. Mostly fades are used between the shots, no hard cuts, and the camera moves. The most significant one starts at 0.36: as long as Williams speaks, the camera slowly tilts from his middle up to his face; when it rests on his face, the voice-over starts to speak again.

time	visual mode	speech mode
		[9 seconds of silence] This petition is in a way about what America is and what it offers to its people, the right to strive, to seek and to find purpose.

		My name is Stanley Tookie Williams.
0.23		I have been residing on San Quentin's death row for 24 years.
0.27		I also have the undesirable antonym [?]
0.32		for being the cofounder of the Crips
0.36	[	which today is a genocidal [?] legacy I regret. During his years in prison Williams has undergone a great change. He has educated himself and he has rejected the violence of his past
0.44		and he has promised to work for the future.
0.50		

0.57		As Williams his life has changed, so has he worked to change the lives of others, especially young people

Although not successful in the end, this video is a pragmatically clear, well produced argumentative narrative. An interpreter can effortlessly summarize the argument. Nevertheless the cross-modal integration is quite complex.

We briefly discuss two phenomena. Again the assumption of the unitary source results in in the activation of a discourse scheme. This unitary source assumption – plausible in the genre – works according to the same principles as in the first fragment, but results in a more complicated solution. In the first fragment we saw a single scope blending in which the structure of the speech mode was projected on the visual mode. This we attributed to (a) a relatively weak structure in the visual mode, (b) a strong structure in the speech mode and (c) a specific temporal incongruence. In this second fragment however we see that:

1. again a unitary source is assumed;
2. but there is not one coordinating structure in one of the modes that is a candidate for single scope blending (at least not anymore after 0.23);
3. because several modes show episodes with a structure that carries forward the narrative;
4. and there is no significant temporal incongruence between the modes.

In the first half minute, the situation seems parallel to fragment 1. But then it becomes clear that the narrative - though dominated by the anonymous voiceover - is also developed by other discourse voices; in the fragment by the telephone speech of Williams as well as by his written statements, and later by several other voices in different modes. Most significant for the cross-modal integration is that there is no temporal incongruence between the speech mode and the visual mode. The voiceover is silent - perhaps silenced - when the written statements of Williams ‘speak’ and when Williams speaks in a telephone statement. The obvious temporal cross-modal coordination during Williams’s telephone statement seems out of the hands of the anonymous voiceover as a narrator. The camera movement is significant here. The camera tilts in way that is coordinated with the structure of the voices, but it is hard to attribute this coordination to ‘the voice-over’.

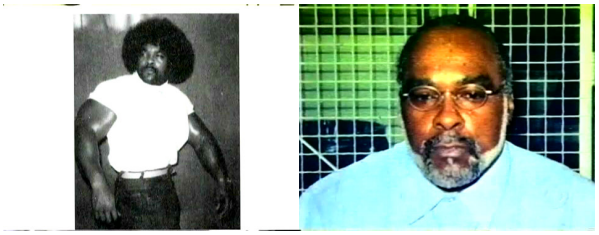
This - we do not attempt a more precise account here - seems to guide the interpreter to restructure the single source assumption from an initial single scope blend into a more complex double scope blend that inherits structural elements from several input spaces. This results in the assumption of an abstract narrator, an organizing principle other than one of the discourse voices.

If this analysis makes sense it illustrates a general principle proposed by Schilperoord & Van den Hoven that claims that conceptual integration is a process that climbs up from simple forms (mirror blending, single scope blending) to more complex forms (double scope blending, multi-scope blending) until an 'equilibrium' is reached, while each 'failure' - an attempt that does not result in a satisfying equilibrium - contributes to the meaning (Schilperoord & Van den Hoven, to appear).

The last phenomenon to be discussed is the most spectacular one to argue the plausibility that a model to account for multimodal discourse processing should assume:

1. immediate processing
2. as well as a strong cross modal connectivity
3. in which the visual mode influences the verbal mode(s).

The example is so obvious that we only need to point at it. Focusing on the speech mode only, the semantics of 'Williams' change' are clearly restricted to a mental and social change. Integration with the visual mode however broadens the semantics in a spectacular way.



One who may doubt whether such a drastic mental change as claimed is possible, is confronted with an undeniable example of a drastic physical change that however is only rhetorically significant if we assume a deep cross-modal integration. So we see how the unitary source assumption - step by step - results in the construction of one coordinating narrator, so one coherent argument, which results in a rhetorically powerful interpretation of a cross-modal parallelism.

Conclusions

In this paper we have used discourse analysis as a method to argue a number of requirements that models for cross-modal integration have to meet. Although in an explorative way, this

shows the strength of such a method that should precede all empirical experimental work. The analysis of two pragmatically straightforward fragments that both are intuitively clear, lead to a number of conclusions that are probably not surprising in a cognitive linguistics environment, but are nevertheless crucial for the question how to approach language in these multimodal texts. We sum up.

1. It is plausible that in the processing of multimodal texts the immediacy principle (all information is used from the moment on that it is available) works across modes.
2. It is plausible that this form of multimodal processing is at least partly in the realm of linguistic connectionism (similar to what is concluded from research on gesture and on multimodal speech processing).
3. A unitary source assumption seems to account for on the one hand perceptual paring (leading to for example a ventriloquism effect), on the other hand to single scope blending and even double scope blending of modal structures, clearly a partially top-down guided, post-perceptual inferential process.
4. Decisions about the blending of modal structures seem to be taken immediately when plausible and lead to specific cross-modal integration processes. When a presumed integration processes does not result in a satisfying solution, a more complex process is attempted which results in a more complex meaning.
5. The complex cross-modal integration processes in which knowledge-sources activated in one mode guide the structuring of other modes, and so on, makes the process of (multimodal) discourse processing a truly holistic process.

References

Bertelson, P., & B. de Gelder (2004). The psychology of multimodal perception. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* Oxford: Oxford University Press. 151-177).

Cohen, A. J. (2001). Music as a source of emotion in film. In Music and Emotion. P. N. Juslin and J. A. Sloboda (eds.), Oxford University Press, Oxford, 249-272.

Coulson, Seana & Todd Oakley (2005), Blending and coded meaning: Literal and figurative meaning in cognitive semantics. In: Journal of Pragmatics 37. 1510–1536

Eemeren, F.H. van & R. Grootendorst (2004). A systematic theory of argumentation: The pragma-dialectical approach. Cambridge: Cambridge University Press.

Fauconnier, G. & M. Turner (2003). *The way we think. Conceptual blending and the mind's hidden complexities*. New York, Basic books.

Hoven, Paul van den (2010a), Peircean semiotics and text linguistic models. In: Chinese semiotic studies 3. Nanjing Normal University press. Nanjing. 201-226.

Hoven Paul van den (2010), The intriguing iconic sign. In: Interstudia 6, Cultural spaces and identities in (inter)action, Alma Mater. Bacău. 37-50.

Jiang Wangqi (2010). East meeting west: holism in discourse analysis. In: Proceedings of The Third International Conference on Multicultural Discourses. Hangzhou.

Lakoff, J. & M. Johnson, Metaphors we live by. Chicago University Press: Chicago.

Schilperoord, J. & P.J. van den Hoven (to appear), Interpreting visual arguments in cartoons. Presented at ICLC 2011, Xi'an.