The Attentional Foundations of Coherence*

Sam Cumming
University of California, Los Angeles

October 1, 2013

The meaning of a complex representation can be greater than the sum of the meanings of its parts. Consider a discourse consisting of segment S_0 followed by segment S_1. If M_0 is the meaning of S_0 and M_1 is the meaning of S_1, then the meaning of the discourse may be represented schematically as

$$M_0 \wedge M_1 \wedge R(M_0, M_1),$$

where R is chosen from a limited inventory[1] of coherence[2] relations.

For instance, R might be the relation that holds between two propositions whenever the second proposition is an explanation of the first. The following two discourses illustrate this choice of coherence relation:

(1) a. John took a train from Paris to Istanbul. He has family there.
b. EXPLANATION(that John took a train from Paris to Istanbul, that John has family in Istanbul)

(2) a. Mary is annoyed. John ate soup last night.
b. EXPLANATION(that Mary is annoyed, that John ate soup last night)

The phenomenon is not limited to linguistic discourse. In the silent film

*Thanks to Gabe Greenberg, Rory Kelly, and the participants in a seminar at UCLA.

[1]Kehler (2002) gives the following list of relations: PARALLEL, CONTRAST, EXEMPLIFICATION, GENERALIZATION, EXCEPTION, ELABORATION, RESULT, EXPLANATION, VIOLATED EXPECTATION, DENIAL OF PREVENTER, OCCASION. Many other taxonomies exist.

[2]Why *coherence* relations? Because a complex representation is judged to be *in*coherent if either (i) the relation between two segments cannot be identified, or (ii) the relation is identifiable but not appropriate. Hobbs gives an example of the latter sort:

(i) John took a train from Paris to Istanbul. He likes spinach.

We sense that an explanation is intended, but can't make sense of the second sentence as an explanation of the first.

Number, Please (1920),[3] analogous coherence relations connect the contents of the individual shots. David Bordwell[4] summarizes a stretch of the action as follows: "Harold Lloyd plays a lovesick boy who's been jilted by his girl. Moping at an amusement park, he sees her arrive with a new beau."


"He shifts to another spot to watch them. When she notices him, she scorns him, and he reacts."


Bordwell uses this sequence to illustrate the ubiquitous technique of *constructive editing,* where spatially non-overlapping shots are used to depict a unified scene. "In this scene... Harold and the couple aren't shown in the same frame. The action is built entirely out of singles of Harold and two-shots of the couple, with an especially emphasized close-up of the girl's snooty reaction."

The content of the second shot (showing the girl with her new beau) *explains* the surprised look from Harold Lloyd's character in the first:

(3) EXPLANATION( , )

[3] https://www.youtube.com/watch?v=pbYPUxNT-qM

[4] http://www.davidbordwell.net/blog/2008/02/04/what-happens-between-shots-happens-between-your-ears/

The interpretation of the second pair of shots exhibits another relation from the standard inventory – that between an action and its effect or *result*:[5]

(4) RESULT( , )

The augmentation of meaning by the application of coherence relations is unexpected on the usual compositional model of semantics, where each part of the meaning of the whole is furnished by some component of the representational complex. This apparent departure from compositionality (whether or not it is *merely* apparent) requires explanation. Moreover, since the same phenomenon crops up in different representational media, a general explanation will be more satisfying than one that bottoms out in medium-specific conventions, such as the lexical semantics of particular words.[6]

In this paper, I suggest that the source of some coherence relations is

[5]These higher-level conceptual relations may be inferable from more basic spatial (and temporal) coherence. In this case, we are given to understand that each pair of shots is connected by Harold Lloyd's line of sight (more on this anon), and from this spatial connection we can use commonsense reasoning to infer causality. However, coherence relations analogous to EXPLANATION may also occur in film in the absence of any spatial connection. For example, the "switch-back" – an editing technique used by D. W. Griffith and others – consists of an insert cut to the psychological cause of a character's behaviour, without any attendant spatial meaning. In *The Salvation Army Lass* (1908), the shot of a character who is hesitating on his way to commit a burglary is intercut with one of his wife still slumped on the ground where he had earlier pushed her aside – and indeed we understand that it is his wife's plight at his own hands that explains his hesitation and eventual abandonment of the planned crime. Though the would-be burglar looks off-screen as he hesitates, this does not necessarily establish a line of sight connecting him to his wife. Persson (2003: 95) speculates that the switchback interpretation competes with the line-of-sight link, and that the former convention would be more familiar to audiences of the period:

> To an early viewer who was familiar with the switchback convention, such a POV connection would probably be weaker and less probable. Here the switchback interpretation, with its duality as both mental image and objective cutaway to another (distant) place in the diegesis, would suffice, without resorting to a glance-target relationship between shots.

[6]Asher and Lascarides (2003) account for the interpretation in (2) by including the augmented meaning in the lexical entry for the verb 'annoy'. As Gabe Greenberg pointed out to me, the lexical account won't even extend to related discourses that omit 'annoy', such as the following:

the attentional system. When we hear that Mary is annoyed, we tend to wonder why. Our attention is naturally occupied with the question, 'Why is Mary annoyed?'. A follow-up statement, such as 'John ate soup last night', is then apt to be taken as answering this implicit question,[7] which means it is understood as providing an *explanation* of Mary's state.

In sum, we derive the extra meaning in (5-b) by interpolating a question (summarizing the orientation of attention at the juncture between the segments) into (5-a):

(5) a. Mary is annoyed. [*Why?*] John ate soup last night.
b. EXPLANATION(that Mary is annoyed, that John ate soup last night)

This works because we treat the meaning of the question as the property of being an explanation of Mary's annoyance (Groenendijk & Stokhof 1988):[8]

(6) $[\![\text{Why is Mary annoyed?}]\!] =$
$\lambda p.\text{EXPLANATION}(\text{that Mary is annoyed}, p)$

This account predicts the same coherence relation – EXPLANATION – no matter how we convey Mary's psychological state in the first segment (for instance, via a comic-strip panel or a film shot). Even if we conceive of the additional meaning as being due to medium-specific conventions, the account purports to identify the common source of those conventions (this will be elaborated in a later section).

Different coherence relations correspond to different interpolated questions, which in turn summarize different attentional states. For example, the RESULT relation would correspond to the question 'What happened as a result?', which presents a state of attention to the consequences of the action or event just described. Note that the attentional states summarized by questions are particular and multiplicitous, while coherence relations correspond to abstract categories of attentional transition.[9] For instance, while both (1) and (2) provide examples of the EXPLANATION relation, the im-

(i) Mary is scowling. John ate soup last night.

[7]So long as it is pronounced with an intonation contour appropriate for such an answer. See Kuppevelt 1995 for details.

[8]Moreover, we treat the total information conveyed by a discourse consisting of a statement S_0 followed by a question-answer sequence Q_0, S_1 as:

(i) $[\![S_0]\!] \wedge [\![S_1]\!] \wedge [\![Q_0]\!]([\![S_1]\!])$

[9]Hobbs' relation of ELABORATION can be seen as a blanket term for a miscellany of question-types (Rohde 2008, Ch.6). For instance, both of the following are elaborations:

plicit questions raised by their respective first segments are distinct. They are 'Why did John take the train from Paris to Istanbul?' and 'Why is Mary annoyed?', respectively.

The best support for the attentional hypothesis comes from film. We can feel confident about the attentional source of coherence relations in film, due to the wealth of data on visual attention and the well-documented origins of editing conventions in film. Indeed, the parallel between editing and attentional transition has been remarked upon since the beginning of theorizing about film.[10] Smith (2006: 62) gives this representative quotation from Ernest Lindgren:

> "[Editing] reproduces the mental process... in which one image follows another as our attention is drawn from this point to that in our surroundings. In so far as the film is photographic and reproduces movement, it can give us a life-like semblance of what we see; in so far as it employs editing, it can exactly reproduce the manner in which we normally see it." (Lindgren 1948: 54)

Film is a representational medium unlike symbolic language, in that it requires minimal prior acquaintance to be understood. The conventions by which filmmakers imbue their representations with meaning are, for the most part, apprehensible on first exposure. This is because they are decoded, in many cases, by the application of inborn mechanisms for making sense of the world, such as our faculties of vision and audition.

The idea that certain film edits are decoded – and the "logic" of the shot sequence recovered – with the help of our native system of visual attention is clearly formulated by Tim Smith (2006: 62).

> When viewing reality our saccadic eye movements present our perceptual system with a succession of views, all of which are presented in response to some form of perceptual inquiry (Hochberg & Brooks [1978]). For example, the perceptual question "What is that man looking at?" is answered by a saccadic eye movement to the target of his gaze. At this point the question is

(i) a. Bill can open John's safe. [*How can he?*] He knows the combination.
b. A young, aspiring politician was arrested in Houston today. [*Who was it? Where did it happen? Why was he arrested?*] John Smith, 34, was nabbed in a law firm while attempting to embezzle funds for his campaign.

[10] Smith (2006: 99) provides a list: D. W. Griffith in Jesionowski 1982: 46; John Huston in Bachmann 1965 and Sweeney 1973; Dmytryk 1986; Katz 1991; Lindgren 1948; Murch 2001; Münsterberg 1970; Pepperman 2004; Reisz & Millar 1953.

> answered by the object now occupying the centre of the viewer's attention. This perceptual question was endogenously answered but a similar question could also be answered by the answer itself capturing attention (i.e. exogenous control). The "snap" of a twig off to your side whilst you walk through the woods elicits involuntary orienting to the source of the sound. The initial "snap" poses the perceptual question, "What made that sound?" which is answered by the eyes being captured by the cause. The same pattern of perceptual question and answering...occurs whilst watching film. The main differences lie in the extent of reorienting and the locus of control: the eyes can never be directed beyond the screen edge and whilst the viewer may want an answer they can never get the answer unless the editor gives it to them. These differences may change the perceptual consequences of a filmic Q&A sequence compared with reality...but in terms of the distribution of attention it is very similar.

Smith discusses the example of a SIGHT LINK between two shots in some detail.[11] This common spatial coherence relation places the content of one shot (known as the "object shot") on the line of sight of a character whose glance is shown in the other shot (known as the "glance shot"), thereby connecting the (usually non-overlapping) spaces disclosed on either side of the cut.[12]

The first two shots of the sequence from *Number, Please* are related by SIGHT LINK:

(7) SIGHT LINK( , )

The (augmented) interpretation of this sequence is that the main action in

[11] An analogous account could be given of psychological relations (e.g. the switchback relation from early cinema – see Gunning 1991, Ch.4) where one shot shows the character thinking and the other gives the content (or possibly just the topic) of the thought.

[12] Note that the camera position in the object shot need not correspond to the orientation of the character's glance, as it does in a true point-of-view shot. While the shots in the second pair from *Number, Please* are of connected glances, neither glance is directly into the camera, and hence the camera is not positioned on the connecting line of sight in either shot.

the second (object) shot takes place in the direction of the glance disclosed in the first shot.

Smith considers the case where the glance shot precedes the object shot, and claims that the glance automatically raises the perceptual question 'What is the character looking at?' in the mind of the viewer.[13] In this case the editor – rather than an autonomous saccade – satisfies the viewer's curiosity, by cutting to the object shot. The viewer, accustomed to having such inquiries answered immediately in ordinary viewing situations, treats the content provided by the editor as the answer to the perceptual question.[14] Since the character couldn't be looking at the content depicted in the object shot unless it lay in the path of their gaze, it follows that the glance shot and the object shot are connected by SIGHT LINK.

[13]He defends this claim at length (2006: 65–7), summarizing as follows (by "deictic cue," he means an element in the scene that initiates a logical or causal connection – a coherence relation – between shots):

> This potential for eyes to first *attract*, by "popping-out" of the visual scene, and then *direct* attention is critical for the use of gaze as a deictic cue in editing. If a viewer is fixating the eyes of an actor, when those eyes suddenly shift and point across the screen, the viewer's attention will be involuntarily pushed in the same direction (Driver et al., 1999; Friesen & Kingstone, 1998). The viewer's eyes will not move, as there is not yet a target for them to move their eyes to, but their attention will covertly shift in the direction of the gaze. This shift in attention combined with the viewer's ability to read intentionality into another person's gaze (Baron-Cohen, 1995) leads the viewer to expect a target for the gaze. [Smith 2006: 67]

[14]Smith (2006: 67) goes into further mechanical detail:

> In real-world vision, the viewer would then use their cued attention to either locate an object in the periphery of their vision or move their head to locate an object out of view. They would then perform a saccadic eye movement to the first object they found that aligned with the gaze. In film, the same projection of the gaze through visual space will occur but it will stop as soon as it reaches the screen edge. If the target of the gaze is found within the screen a saccadic eye movement will be initiated.... If no valid target exists the editor will have to provide one by cutting to the point/object shot. The object depicted in the point/object shot can either be located along the path of the actor's gaze, requiring a saccade to fixate..., or be collocated with the viewer's current point of fixation.... In the latter case no saccadic eye movement is required but attention will still be captured by the sudden onset of the expected object.

(8)  [*What is he looking at?*] 

Smith's account of the *augmented meaning* in a SIGHT LINK sequence as a perceptual *question* corresponding to a state of visual *attention* matches the template of the attentional account of coherence previously sketched. Indeed, it was the inspiration for that account. My plan for this paper is to extend the account of SIGHT LINK in film to other cases of coherence.

The next stage of the paper rounds out the presentation of the positive view. I first motivate the project of addressing the psychological foundations of coherence and discuss some of the alternatives to the attentional account. I then give an account of attentional focus on which it may be expressed as a question, and use this to show how the match between attention and coherence might be assayed empirically. Finally, I address some of the complications on the route from attention to augmented meaning. For the remainder of the paper, I discuss the limits of the attentional account, considering examples of coherence relations that are not attentional in origin, and conclude with a survey of related work on attention in semantics.

1 The Psychology of Coherence

Once it is observed that the meanings of complex representations are augmented by coherence relations, the empirical project proceeds in two general directions. One course relates the choice of coherence relation to other features of the representation. For instance, its effect on how anaphora is resolved, how reference time progresses, what kinds of syntactic movement and ellipsis are possible, which discourse particles are barred, etc.[15] The second descriptive project addresses the range and organization of the coherence relations themselves.[16] It is usually assumed that the available relations must be restricted if the theory is to make predictions (e.g., about which sequences are coherent – Knott & Dale 1994).

[15] As an illustration, the EXPLANATION relation privileges coreference with the "causally implicated referent" of the initial sentence, halts (or reverses) the advance of temporal reference, eschews syntactic parallelism in VP-ellipsis, and is inconsistent with the conjunction 'and' (among others).

[16] Another pressing issue, addressed in Hobbs et. al. 1993 and Asher & Lascarides 2003, concerns how agents coordinate on a particular choice of coherence relation.

Many inventories of coherence relations have been proposed to date.[17] All, so far, have been based on the theorist's intuitions and judgment.[18] Once a list is produced by this method, the theorist is frequently prompted to reflect on its underlying principle. As Kehler puts it:

> We expect that there are fundamental cognitive principles at work which will serve both to constrain the set of possible relations, and to provide an explanation for why a particular set of relations is to be preferred to one containing more, fewer, or different relations. (Kehler 2002: 25)

The focus on the descriptive enterprise means that there is normally not the scope for a full-scale investigation of the foundations of coherence. Instead, many different speculative hypotheses have been advanced, all of which trace the source of coherence relations to some aspect of agent psychology or rationality.

Take, for instance, this early proposal from Jerry Hobbs:

> It is tempting to speculate that these coherence relations are instantiations in discourse comprehension of more general principles of coherence that we apply in attempting to make sense out of the world we find ourselves in, principles that rest ultimately on some notion of cognitive economy. (Hobbs 1985: 23)

He has in mind the tendency towards *consilience,* or the use of the same hypothesis to account for a multiplicity of data:[19]

> We get a simpler theory of the world if we can minimize the number of entities by identifying apparently distinct entities as different aspects of the same thing. Just as when we see two parts of a branch of a tree occluded in the middle and assume that they are parts of the same branch, so in the expansion relations we assume that two segments of text are making roughly the same kind of assertion about the same entities or classes of entities.

[17] Kehler (2002) lists the following: "Halliday and Hasan 1976, Hobbs 1979, Longacre 1983, Mann and Thompson 1987, Polanyi 1988, Hobbs 1990, inter alia; see Hovy (1990) for a compendium of over 350 relations that have been proposed in the literature." To this we must add Knott 1996.

[18] See Knott 1996 for extensive ruminations on this method.

[19] McCloud (1993) makes a similar connection between coherence and *closure* in gestalt psychology.

> When we hear a loud crash and the lights go out, we are apt to assume that one event has happened rather than two, by hypothesizing a causal relation. (Hobbs 1985: 23)

Hobbs also cites with approval David Hume's distillation of three basic principles by which we associate ideas – Resemblance, Contiguity in time or space, and Cause-Effect. Both Hobbs and Kehler structure their taxonomies of coherence relations using Hume's principles as a guide. However, the appeal to Humean associationist psychology is, I think, in tension with the rational appeal to consilience. While it is true that contiguity in space (say) occasionally counts as a consideration in favour of unification – as in Hobbs' own example of the contiguous sections of tree branch – the more usual case of association concerns entities known to be distinct.

Indeed, such associations can *disconfirm* the hypothesis that two entities are the same. For instance, if two manifestations, not contiguous in space, exhibit close enough contiguity in time, we can conclude that they belong to distinct entities. Similarly, resemblance in projected retinal image can count against unification if in each case the object is viewed at a different distance from the eye, or under different lighting conditions. And, strictly speaking, causes are distinct from their effects, not identical to them.[20]

The considerations that lead to a judgment of unification are subtle, variable, and sensitive to what we know. It seems unlikely that such judgments routinely appeal to Hume's principles, even heuristically.

Even if they cannot be said to follow from the rational principle of consilience, Hume's associative principles live on in psychology. As Kahneman (2011) writes, "Our concept of association has changed radically since Hume's days, but his three principles still provide a good start." According to him, "Psychologists think of ideas as nodes in a vast network, called associative memory, in which each idea is linked to many others." Perhaps those who endorse Hume's account of the fundamental cognitive basis of coherence could appeal to these structuring principles of associative memory.[21]

[20] Though admittedly being so related allows us to explain both with the resources for explaining one.

[21] While Kehler uses Hume's principles to structure his taxonomy, he seems not to have associative memory in mind, but instead something tantalizingly close to the attentional theory:

> It is likely that we will gain a greater understanding of the theoretical status of discourse topichood as we understand more about the larger questions concerning the factors that determine coherence. My personal suspicion, however, is that the discovery of these principles will only go so far using the semanticist's tools of possible worlds and logic-driven inference. As in the

Unfortunately, an exhaustive store of potential associations is problematic from a computational perspective, as Randy Gallistel has previously insisted:

> It is a disadvantage of associative theories of memory that they presuppose an associative path (not necessarily direct) between a memory and any other memory capable of evoking it in some context or given some task. In a system with many memories to be accessed for many different purposes in many different contexts, this leads to an explosive proliferation of associative links.... Even if we abandon the assumption that each memory is directly accessible from each other, it seems clear that in any large memory net, the preponderance of the system must be given over to the linkages between records (the associative bonds) rather than to the records themselves.
>
> The system may resort to some kind of hierarchical organization ("chunking of memories") to keep the required number of links within bounds, but it is not clear that this could overcome the dilemma inherent in a scheme that attempts to link records a priori before a particular connection is required for some memory-dependent output. The dilemma is that any scheme to economize on linkages (on what is linked to what) limits the use that may subsequently be made of the recorded data... (Gallistel 1990: 541–2)

Though I won't develop this idea further, the associative connections that Kahneman is talking about (and which Gallistel views as the brute motivation behind associative psychology) might in many cases correspond to transitions mediated by the attentional system (eliminating the need to store them in a sector of memory). Instead of grounding coherence relations in associative connections, there is the potential to ground both (or at least parts of both) in the mechanisms of attention.

> domain of vision, many cognitive factors come into play that bear on the way in which we perceive the world as coherent – our attentional mechanisms, statistically driven expectations, the salience we accord different properties when judging similarity, and so forth. (Kehler 2004)

2 Attention and Questions

Previous authors have drawn a connection between coherence relations and implicit questions (e.g., Roberts 1996: 50), and even probed it experimentally (Rohde 2008, Ch.6). While it is often possible to match implicit questions to individual coherence relations – for instance, the interpolated question 'Why?' mimics the content and argument order of the EXPLANATION relation – some differences between the accounts must be acknowledged.

(9) a. John took a train from Paris to Istanbul. [*Why?*] He has family there.
b. EXPLANATION(that John took a train from Paris to Istanbul, that John has family in Istanbul)

Obviously, implicit questions are not uttered, whereas coherence relations are sometimes (and, in the case of certain relations, must be) signalled by a discourse particle (this point will be revisited later on).[22] Moreover, the two accounts offer different approaches to the structuring of discourse. For example, Kuppevelt admits arrangements in which a single statement (called a "feeder") gives rise to multiple questions (1995: 122), or on which the answer to one question acts as the feeder for another (1995: 130). But what seem like natural analyses on a question-based account correspond to controversial non-tree-like structures on relation-based accounts (Hobbs 1990; Webber et al. 2003; Wolf & Gibson 2006).[23]

[22] In some cases the discourse particle can coexist with the question (and it is perhaps worthy of note that in some languages, for instance Modern Greek, the question word and the discourse particle are identical). For instance:

(i) a. Mary is annoyed. Why? Because John ate soup last night.
b. John ate soup last night. What happened as a result? As a result, Mary threw a fit.
c. Gareth grew up during the seventies. So what? So he loves disco music.

In other cases, there is no congruent question:

(ii) Edith grew up during the seventies, but she hates disco music.
a. ?? Edith grew up during the seventies. But what? But she hates disco music.
b. #Edith grew up during the seventies. Does she like disco music? But she hates disco music.

Note (ii-a) would be acceptable if the speakers alternated and the first speaker nonverbally indicated some qualification at the end of the first sentence.

[23] The PARALLEL structure receives a more explanatory treatment on the implicit question account. Two segments related by PARALLEL are best thought of as partial answers

The literature on implicit questions in discourse addresses how a question might be recovered from the content and intonation of its *answer* (what would be the second relatum of the coherence relation). For instance, the placement of focal stress (represented by capitals) enables us to recover the original position of the question word:

(10) a. [Who hit Bill?]
HARRY hit Bill.
b. [Who did Harry hit?]
Harry hit BILL.

However, recent experimental work has shown that material in the initial segment (first relatum) can also influence the audience's expectations about coherence, before the answer (second relatum) has even been given. For instance, upon seeing an implicit causality verb (such as 'annoy'), audiences treat EXPLANATION as the most likely relation to bind the sentence they are currently reading to the one to follow. By contrast, upon seeing a verb describing a transfer of possession, audiences expect the next sentence to be linked by OCCASION (Rohde & Horton 2010).

The attentional theory of coherence is in a position to explain these results, since it posits a relationship between the audience's *expectations about coherence* and their *attentional state* at the juncture between segments.[24] For instance, if implicit causality verbs shift audience expectations towards EXPLANATION, we might also expect them to shift attentional focus to the question 'Why?'.[25]

It follows that the attentional theory is testable. As mentioned, experimental linguists have developed various means of probing online expectations of coherence relations. Meanwhile, psychologists have ways of determining what is in the focus of attention.[26] Crucially, the results of the two

to the same question (or else as complete answers to different subquestions of the same question). This contrasts with other relations that are associated with an *intervening* question – one raised by the initial segment and answered by the subsequent one. See Büring & Kehler 2007.

[24]Kuppevelt 1995 also touches on the relationship between the initial segments he calls feeders and the implicit questions they raise. He points out that indeterminate expressions such as 'someone', 'sometime', 'somewhere', etc., provide a locus for further questioning – 'who?', 'when?', 'where?' (1995: 120).

[25]This illustration of the relationship between attention and expectations of coherence is too simplistic on a number of counts. Complications will be addressed later in this section, and in the next.

[26]While the lion's share of this research has been conducted on visual attention, some parallel results have been obtained in discourse comprehension. For instance, Sanford

sorts of test are comparable, since both coherence relations and attentional focus may be modelled as *questions*.

What captures our attention depends partly on the stimuli in our environment and partly on what we are able to ignore.[27] It turns out that even salient and distinctive stimuli (such as oncoming cyclists, or men in gorilla suits) can fail to capture attention if the attentional system is not tasked with finding them. Attention is versatile, and is capable of locking on to particular frequency bands, levels of depth, colours, and shapes; it can focus on stimuli that are moving, as well as those that are not. If a stimulus falls outside the focus of attention (often called the "attentional set"), it is likely to be successfully ignored.[28]

We can thus isolate two components of attentional state: the stimulus that has captured attention (and thus the object or region that attention is oriented towards), and the sort of thing that the attentional system is presently tasked with finding (what I have been calling attentional *focus* – modelled by attentional set).[29] It is the latter that is of most relevance to the attentional theory of coherence.

Attentional set is usually inferred from (or operationalized as) the set of entities, characterized by some property F, that are relevant to the task

2002 explores parallels within the change-detection paradigm (Simons & Levin 1997).

[27] Attentional capture can be *explicit* – so that the subject becomes aware of the stimulus – but it can also occur without the conscious awareness of the subject, in which case it must be inferred from eye-movement or an effect on reaction time (Simons 2000). Attentional capture is *covert,* in the sense that it doesn't necessarily occur at the fixation point of the eyes. However, it often heralds eye-movement to that position.

[28] Here is Simons summarizing the results from a particular experimental paradigm:

> In [the pre-cuing] paradigm, the attentional set of the observer plays a critical mediating role in attentional capture, even by abrupt onsets. For example, color pre-cues only capture attention if subjects are searching for a color target. If the attentional set is for dynamic stimuli (e.g. a motion or late-onset target), only dynamic pre-cues capture attention and if the attentional set is for static stimuli, only static pre-cues capture attention. Furthermore, when observers are searching for a specific feature value (e.g. green), only pre-cues of the same value (e.g. green but not red) capture attention. In general, attention can be captured by any singleton when observers are in a singleton search mode, but if they are searching for a particular feature value, only cues with the same value will capture attention. (Simons 2000: 149)

[29] It is reasonable to treat attentional set as a partial characterization of the attentional system itself (some are even willing to operationalize attention *as* task relevance – Summerfield & Egner 2009), though there are those who treat it as an external cognitive manipulation of attentional capture.

the subject is engaged on. For instance, if the subject is told to determine whether the uniquely coloured item in a visual array is an 'E' or an 'H', then their attentional set is assumed to be the set of uniquely coloured items.

Note that any property F used to specify an attentional set can equally be used to specify a question:

(11) ATTENTIONAL SET: $\{x \mid Fx\}$
QUESTION: 'What is F?'

I use these notions interchangeably to characterize attentional focus.

There is a great deal of evidence supporting the influence of attentional focus on attentional capture (Simons 2000; Ruz & Lupiáñez 2002). This means we can often infer the property (or question) in attentional focus from the stimuli that capture attention.[30] On the attentional theory of coherence, attentional capture should additionally provide evidence of the *implicit question* intervening between segments.

Here is a simple example of an experiment (at this stage only hypothetical) that uses attentional capture to test the latter. Suppose we have a film shot that contains both a dynamic (moving) and a static stimulus. Which stimulus captures attention in this shot depends on the viewer's attentional set, since attention can focus on stimuli of either type. Now, we know we can manipulate attentional set by setting the subject a task that requires attention to a particular property. But we should be able to achieve a similar effect with a visual stimulus. For instance, we might suppose that a close-up of a person with a rigid off-screen gaze will focus the viewer's attention on static objects, while a close-up of a person with a gaze moving in smooth pursuit of some object will focus their attention on dynamic objects. This hypothesis could be tested experimentally by cutting from either gaze shot to the shot containing the dynamic and static stimuli, and seeing which one captured attention (e.g., using an eye-tracker). The result will give us fine-grained information about the visual question implicitly raised by the initial shot (is it 'What is he looking at?', or 'What *moving thing* is he looking at?'?) and hence the coherence relation linking the shots. The attentional theory thus offers a novel means of testing coherence, one that does not rely on the subjective judgments of expert (and potentially theory-driven) linguists.

A complication must be addressed before we move on. As noted earlier, Kuppevelt observes that a single sentence can raise *more than one* question. Here is an example:

[30] Other empirical tests, such as brain imaging, may also be used (Datta & DeYoe 2009).

(12) a. John is ill.
b. How long has he been ill?
What does he suffer from?

How are we to explain this on the attentional account? We cannot say, for instance, that the segment places a particular *sequence* of questions in focus, since both discourses below are perfectly acceptable, though they address the questions raised in a different order.

(13) a. John is ill. [*How long has he been ill?*] He has been sick for three days now. [*What does he suffer from?*] They think it's bronchitis.
b. John is ill. [*What does he suffer from?*] They think it's bronchitis. [*How long has he been ill?*] He has been sick for three days now.

Rather, we must think of the attentional system as simultaneously[31] engaged by several issues, which can be addressed in any order (the content and structure of the answers will tell us *which* order). We must therefore model attentional focus as a *set* of questions (or set of attentional sets).

There might be reason to adopt a more specific model that imposes some order, or even metric, over the questions in the set.[32] For instance, we could follow Philipp Koralus (2013), who models attentional states as assignments of degrees of *sensitivity* to questions. In signal detection theory, sensitivity is a measurement of a system's ability to discriminate between cases (e.g., the presence or absence of a stimulus).[33] Since it is well established that attentional focus boosts the subject's ability to discriminate, it might seem natural to measure the attention devoted to a question by the system's sensitivity to which answer is the case.[34]

Unfortunately, this detailed way of modelling attention doesn't apply to the scenario we are interested in: the interpretation of complex representations. In that case, we rely on an editor or author to provide the answers to the questions in attentional focus. It is up to them whether to answer immediately, to delay the answer, or even to withhold it entirely. As a result, our sensitivity to the answer is decoupled from the degree to which our

[31] Perhaps by rapidly switching among them. I won't address this issue here.

[32] One such reason would be to predict the uneven expectations of coherence relations described by Kehler and colleagues.

[33] The sensitivity index is estimated from the subject's hit rate and false-alarm rate in performing the task.

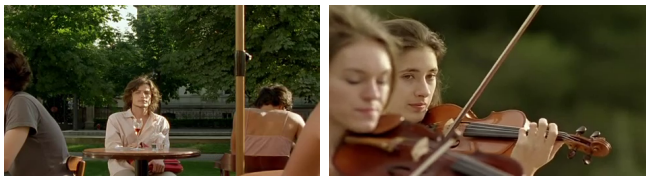
[34] Note that Koralus's own account differs from this straightforward one.

attention is focused on it.

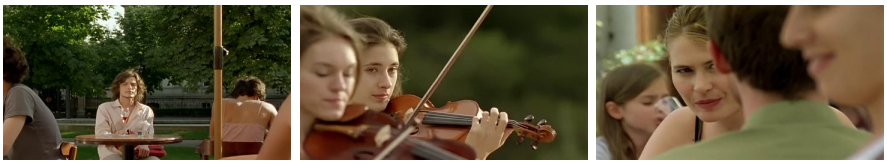
3 From Attention to Meaning

A coherence relation (or the corresponding implicit question) is a piece of content that must be coordinated on by communicating agents. Suppose, to begin with, that the implicit question connecting two segments always summarizes the audience's attentional state after processing the first segment. The representation's author and the audience could then coordinate as follows. The audience would simply interpret each segment as the answer to an implicit question summarizing his or her state of attention at the beginning of the segment.[35] The author could anticipate this interpretation by modelling the audience's attentional state – something made possible by a shared psychological endowment and overlapping knowledge.

Sometimes, however, the author deliberately foils the audience's expectations. For instance, José Luis Guerin, in his film *In the City of Sylvia* (2007), cuts from a shot of his protagonist's glance to a shot of two female violinists, one of them in sharp focus:



We are initially inclined to relate these shots with SIGHT LINK.[36] In doing so, however, we have fallen victim to a ruse. We soon discover that the editor has inserted the shot of the violinists between the glance shot and the true object shot of the SIGHT LINK:



[35] In a sense, the audience treats the representation (or its author) as an oracle, accepting its pronouncements in place of the results of the usual process of inquiry.

[36] Indeed, we tend to see the second shot as taken from the *point-of-view* of the protagonist. The particularized focus tips us off.

Decoding this sequence demands more of the audience than the procedure sketched above. The audience follows that procedure at first – treating the content of the second shot as an answer to the implicit question 'What is he looking at?' – but then proceeds to revise the interpretation once the third shot is screened. There is no provision for this sort of revision in our initial model of coordination; indeed, this is a case where the implicit question between shots one and two doesn't match the audience's attentional focus at that juncture.

We might suppose that the audience's attention is *divided* at the moment of the cut. In addition to 'What is he looking at?', the audience is also mildly wondering 'Where is that music coming from?'. Rather than answering both questions at once (as we think at first), the second shot only answers the second of the two. It is still true that the shots are related by *some* question the audience was attending to at the cut.

However, we can challenge even this weakened proposal.[37] In *Once Upon a Time in the West* (1968), the scene where the Claudia Cardinale character arrives at the farm opens with a shot of her troubled glance. Before the cut to the subsequent (point-of-view) object shot, the audience can already guess its contents: the aftermath of the massacre shown earlier in the film. Though the audience (in a sense) already knows the answer to the question 'What is she looking at?', they still expect the next shot to answer that question (indeed, the expectation, combined with the knowledge, creates a queasy suspense). Thus the implicit question connecting two segments sometimes doesn't belong to those the audience is still wondering about.

There are various ways one might respond to this argument.[38] Still, in my opinion it hits its mark. The implicit question connecting two segments need not be part of the audience's actual attentional state. Minimally, we should require it to correspond to an attentional state that is *intelligible* in the context, making it possible for the communicating parties to coordinate on it. The audience might think, 'Even though I already know what Claudia Cardinale's character is looking at, I can imagine someone who didn't know wondering about it at this point. Perhaps that's what we'll be shown next.'

The version of the attentional theory I wish to endorse says that a certain subset of the coherence relations originate from the attentional system, in

[37] The following example is due to Rory Kelly (p.c.).

[38] Though we know in vague terms what we will see after the cut, we might still be interested in the detail. Moreover, in the case of a point-of-view edit, we might replace the question 'What is she looking at?', with 'What is it like to be her – seeing through her eyes – right now?' whose answer demands such additional detail. Nevertheless, we can imagine further counterexamples that would thwart such defensive manoeuvres.

the sense that they correspond to natural attentional transitions between segments. Since both communicating parties have a grasp on the range of attentional states that are natural in various situations, they possess the basis on which to coordinate on a particular relation for a particular pair of segments. Instead of being read directly from their own attentional state, an audience's expectations of coherence will be calculated from the speaker's signals. For instance: 'Typically, one follows a glance shot with a shot that would satisfy the curiosity of one who did not already know, as I do, the target of that glance. Most likely that will happen again here.' This sort of extrapolation is the hallmark of convention, and ultimately it is the foundations of "natural" conventions of discourse and cinema that I intend the attentional theory to clarify.[39]

As a final complication, it's worth mentioning that sometimes we apprehend an attentional transition even when the segments are presented out of their natural order. For instance, the object shot can occur *before* the glance shot in a SIGHT LINK series. Occasionally, the augmented interpretation will admit an alternative analysis in the form of a question-and-answer sequence matching the reversed order. For example, Gabe Greenberg, Rory Kelly and I constructed a clip in which the shot of a chessboard in mid-game is followed by a close-up of a person looking intently at something off-screen. It is natural to ask, upon seeing a chessboard in mid-game, 'Who is playing this game?'. But if we suppose the shots are linked by this question, we can infer the SIGHT LINK content, given that one playing a game of chess is likely to be looking at the board.

In other cases, however, the subsequent glance shot provides the only evidence in favour of the SIGHT LINK, and that connection is not expected until the glance appears. In those cases, the audience needs to recognize that the answer has preceded the question it is answering. While the coherence relation still has its basis in attention, this sort of case, strictly speaking, does not conform to a "natural attentional transition" between segments, due to its reversal of the natural order.

The situation is complicated by the fact that there are also clear influences of order on coherence. For some coherence relations, the reverse order is not possible. Explanations (unless signalled by an explicit 'because') always follow their explanandum, and it would be nice to account for this using the natural flow of attention (which seeks out explanations, rather than the things to be explained by them). Furthermore, Gabe Greenberg, Elsi Kaiser, Rory Kelly and I have discovered that while the object and

[39] I follow (quite closely!) in the footsteps of Tim Smith (2006) here.

glance shots in a SIGHT LINK can occur in either order, there is an influence of order on the more specialized "point-of-view" (POV) interpretation, which can be eliminated by placing the object shot first.[40]

It is tempting to treat some of this complexity as historical in origin. Discourse and film are both conventional representational systems, and thus the range of coherence relations associated with each may be to some extent arbitrary, the outcome of historical forces and accidents.[41] In the case of film, however, historical explanation only goes so far, and in particular cannot explain how it is that the conventions of film are renewed for each fresh generation of viewers. Historical forces may account for the popularity of parallel editing during the early years of American cinema – in contrast with, for instance, cut-ins (Gunning 1991: 80) – but cannot explain the intelligibility of parallel editing and cut-ins for audiences witnessing them for the first time.

4 The Scope of Attentional Coherence

There are coherence relations that do not comfortably inhabit the purview of the attentional account. One example might be Hobbs' relation of VIOLATED EXPECTATION, of which he offers the following cases:

(14) a. This paper is weak, but interesting.
b. We are in favour of a democratic republic as the best form of the state for the proletariat under capitalism; but we have no right to forget that wage slavery is the lot of the people even in the most democratic bourgeois republic.

Hobbs' definition of VIOLATED EXPECTATION specifies that the second segment contradicts some commonsense entailment of the first. For instance, in the case of (14-a), this would be the implication that the paper is not publishable.

Clinging to the attentional account, one might suggest that upon hearing that the paper is weak, the audience wonders 'Should it be rejected?'. Following up with the claim that it is interesting, the author secures coherence

[40] We have identified a number of factors that influence the POV interpretation. Solely by changing the order of the shots, one can, in the right circumstances, eliminate the POV interpretation in favour of a mere SIGHT LINK. In certain circumstances, however, it is also possible to achieve the POV interpretation even though the object shot precedes the glance shot.

[41] See Greenberg 2011; Lepore & Stone MS.

by offering a partial (though vacillating) answer to that question.

Note that the account above doesn't make the mistaken prediction that the discourse in (15-a) is coherent, as once we add the intervening question, it remains rocky:

(15) a. This paper is weak. It is interesting.
b. This paper is weak. [*Should it be rejected?*] It is interesting.

We can smooth things out by adding an indication of demurral, which begins to explain why VIOLATED EXPECTATION must be accompanied by a discourse particle with negative polarity (such as 'but'):

(16) This paper is weak. Should it be rejected? Well, it is interesting.

The attentional theory was proposed to explain the surprising *unmarked* augmentation of meaning in complex representations. Perhaps we should not think of it as responsible for relations that require explicit marking. The correct origin story for these is going to be at least partially etymological (e.g., involving the journey from a productive phrase to an idiom).

Though it occurs unmarked, I'm also inclined to exclude Hobbs' OCCASION relation from the ranks of the attentional coherence relations. This relation connects segments bridging a change of state (Hobbs 1990), and presents those segments in iconic temporal order (in this case, too, order has a pronounced semantic effect). For example:

(17) a. Add chocolate mixture. Bake for an hour. [From a recipe for chocolate lava cake]
b. Bake for an hour. Add chocolate mixture. [From a recipe for a cake iced with chocolate]

While it is not too difficult to find a question with equivalent meaning ('What do I do next?'), the attentional account lacks explanatory power, as it fails to privilege iconic order. It would be just as natural to attend to the preceding event ('What do I do before that?'), but there is no coherence relation corresponding to this transition. If attention does play a role in the origins of OCCASION, then iconicity plays an equal one.

It is salutary to consider the representation of temporal progression in film, in this context. For film, the representational mechanism is transparently iconic: the progression of time in the shot represents the progression of time in the depicted scene, typically at a 1:1 ratio. This convention is apprehended instantly by viewers. It is not necessary to invoke attentional transitions, or implicit questions, between the frames of a film to account

for the temporal coherence between them.[42]

Certain spatial constraints also contribute to the interpretation of film without appearing to be attentional in origin (though they, like other forms of iconicity in film, most likely have a perceptual source). Gabe Greenberg, Rory Kelly and I found that viewers will use the assumption that a piece of action (such as a car chase or a conversation) is being shown from the same side[43] in different shots to single out one interpretation of a clip from alternatives that are otherwise equally plausible. This assumption, viewed as a spatial coherence relation similar to SIGHT LINK, cannot be cashed out in terms of a question or an attentional state. It is, however, iconic, in the sense that it involves the use of gross screen direction to represent a constant direction in the depicted scene.

Another example of the iconic use of editing is the representational use of the pacing of cuts. The most common technique is to represent an approaching climax by reducing the time between cuts. According to Gunning, this technique first appears in Griffith's film *The Call of the Wild* (1908), where the diminishing shot lengths in a chase scene are used to reinforce the message that the pursuers are gaining on the pursued.[44]

5 Related Work

The present work is preceded by several proposals to link notions of attention to semantics. Most notably, Barbara Grosz and Candace Sidner (1986) also offer an account of the influence of attention on coherence. However, they mean something different from what I do by 'attention', 'coherence', and even 'account'.[45] Still, formal semantic proposals in this tradition, such as those of Bittner (2001; 2007; 2012) and Stone, Stojnic and Lepore (MS), count as allied attempts to regiment semantic conventions with attentional origins.

[42] See Greenberg 2011 for a formal approach to iconic representation.

[43] This is a simplified reference to the 180°-rule used by filmmakers in constructive editing.

[44] Since the decreasing lead is actually depicted in the shots of this sequence, this example doesn't conclusively demonstrate that the pacing of cuts carries meaning. However, if Griffith had cut from the pursued to the pursuer without showing both in the same shot, the cutting rhythm would presumably have conveyed this meaning on its own.

[45] Their attentional focus consists of entities that have been recently mentioned (and hence are candidates for pronominal reference), rather than questions the attentional system is tasked with answering. Their notion of coherence is rational (as in means-end coherence), rather than semantic. Their account is a computational theory of discourse processing, while mine is a foundational theory of meaning.

The literature on information structure and implicit questions, some of whose exponents I have already discussed (Kuppevelt and Roberts), is also seminal to the present account. The most closely related work in this category is Marta Abrusán's (2011) proposal to connect presupposition and the attentional periphery.[46]

References

Abrusán, M. (2011). Predicting the presuppositions of soft triggers. *Linguistics and Philosophy*, 34:491–535.

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.

Bittner, M. (2001). Surface composition as bridging. *Journal of Semantics*, 18:127–177.

Bittner, M. (2007). Online update: Temporal, modal, and de se anaphora in polysynthetic discourse. In Barker, C. and Jacobson, P., editors, *Direct Compositionality*, pages 363–404. Oxford University Press.

Bittner, M. (2012). Perspectival discourse referents for indexicals. In Green, H., editor, *Proceedings of SULA*.

Büring, D. and Kehler, A. (2007). Be bound or be disjoint! In *Proceedings of NELS*.

Datta, R. and DeYoe, E. (2009). I know where you are secretly attending! The topography of human visual attention revealed with fMRI. *Vision Research*, 49:1037–1044.

Greenberg, G. (2011). *The Semiotic Spectrum*. PhD thesis, Rutgers.

Groenendijk, J. and Stokhof, M. (1988). Type-shifting rules and the semantics of interrogatives. In Chierchia, G., Partee, B., and Turner, R., editors, *Properties, Types and Meaning*, volume 39 of *Studies in Linguistics and Philosophy*, pages 21–68. Springer Netherlands.

Grosz, B. and Sidner, C. (1986). Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

[46]This was brought to my attention by Nicholas Asher (p.c.).

Gunning, T. (1991). *D. W. Griffith and the Origins of American Narrative Film: The Early Years at Biograph*. University of Illinois Press, Urbana and Chicago, IL.

Hobbs, J., Stickel, M., Martin, P., and Edwards, D. (1993). Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142.

Hochberg, J. and Brooks, V. (1978). Film cutting and visual momentum. In J. W. Senders, D. F. F. and Monty, R. A., editors, *Eye Movements and the Higher Psychological Functions*, pages 293–317. Lawrence Erlbaum, Hillsdale, NJ.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Strauss and Giroux, New York.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. CSLI, Palo Alto.

Knott, A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, University of Edinburgh.

Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Koralus, P. (2013). The erotetic theory of attention: Questions, focus, and distraction. *Mind & Language*.

McCloud, S. (1993). *Understanding Comics*. Tundra.

Persson, P. (2003). *Understanding Cinema: A Psychological Theory of Moving Imagery*. Cambridge University Press, Cambridge.

Rohde, H. (2008). *Coherence-Driven Effects in Sentence and Discourse Processing*. PhD thesis, University of California, San Diego.

Rohde, H. and Horton, W. (2010). Why or what next? Eye movements reveal expectations about discourse direction. Ms.

Ruz, M. and Lupiánez, J. (2002). A review of attentional capture: On its automaticity and sensitivity to endogenous control. *Psicológica*, 23:283–309.

Sanford, A. J. (2002). Context, attention and depth of processing during interpretation. *Mind & Language*, 17:188–206.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, 4(4):147–155.

Simons, D. J. and Levin, D. (1997). Change blindness. *Trends in Cognitive Science*, 1:261–267.

Smith, T. J. (2006). *An Attentional Theory of Continuity Editing*. PhD thesis, University of Edinburgh.

Summerfield, C. and Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Science*, 13(9):403–409.

van Kuppevelt, J. (1995). Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1):109–147.

Webber, B., Stone, M., Joshi, A., and Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–588.

Wolf, F. and Gibson, E. (2006). *Coherence in Natural Language: Data Stuctures and Applications*. MIT Press, Cambridge, MA.