Discourse, Context & Media 44 (2021) 100544

Contents lists available at ScienceDirect

Discourse, Context & Media

journal homepage: www.elsevier.com/locate/dcm

The impact of multimodal cohesion on attention and interpretation in film

Chiao-I Tseng [a,*], Jochen Laubrock [b,c], John A. Bateman [a]

[a] University of Bremen, Germany
[b] University of Potsdam, Germany
[c] Brandenburg Medical School Theodor Fontane, Germany

ARTICLE INFO

Article history:
Received 10 March 2021
Received in revised form 10 August 2021
Accepted 8 September 2021

Keywords:
Film
Cohesion
Discourse semantics
Multimodality
Eye-tracking
Attention

ABSTRACT

This article presents results of an exploratory investigation combining multimodal cohesion analysis and eye-tracking studies. Multimodal cohesion, as a tool of multimodal discourse analysis, goes beyond linguistic cohesive mechanisms to enable the construction of cross-modal discourse structures that systematically relate technical details of audio, visual and verbal modalities. Patterns of multimodal cohesion from these discourse structures were used to design eye-tracking experiments and questionnaires in order to empirically investigate how auditory and visual cohesive cues affect attention and comprehension. We argue that the cross-modal structures of cohesion revealed by our method offer a strong methodology for addressing empirical questions concerning viewers' comprehension of narrative settings and the comparative salience of visual, verbal and audio cues. Analyses are presented of the beginning of Hitchcock's *The Birds* (1963) and a sketch from *Monty Python* filmed in 1971. Our approach balances the narrative-based issue of how narrative elements in film guide meaning interpretation and the recipient-based question of where a film viewer's attention is directed during viewing and how this affects comprehension.

1. Introduction: from verbal to multimodal cohesion

The analytic tool of cohesion has long been established as a way of characterising aspects of the texture of verbal text (Halliday and Hasan, 1976). Cohesion rests on the observation that 'repetitions' and 're-occurrences' of linguistic patterns appear to be functional for the way in which a text 'holds itself together' as a unit of communication. Although all verbal texts exhibit cohesion of various kinds, associations of cohesion with degrees of comprehension or intelligibility have proved illusive. Correlations between texture and communicative situations, text types, and genres have been investigated both in analyses of particular texts and in corpus studies (e.g., Flowerdew and Mahlberg, 2009). Variations have been observed in relation to the degree of situational-binding required in texts (i.e., spoken *vs.* written, linear *vs.* non-linear texts: e.g., Tanskanen, 2006; Hoffmann, 2012; Schubert, 2017), and patterns of co-reference chains and anaphora have received particularly close attention (cf. Sukthanker et al., 2020). The relationship between cohesion and discourse coherence is evidently complex and several authors have attempted to refine the notion of coherence to support tighter characterizations of the phenomena. Of particular relevance below will be Martin's (1992: Chapter 3) development of a functionally organised discourse semantics for verbal texts. This account sees cohesion as a set of communicative resources for presenting and following discourse referents across any text (focusing particularly on people, places and things), and for classifying links between those elements in terms of specified presentational and tracking strategies.

Phenomena of 'repetition' and 'redundancy' have also long received attention in studies of texts that draw on multiple forms of expression, most commonly verbal language and images but also, for example in film studies, in re-occurrences of musical motifs, particular framing techniques, visual motifs, and so on (cf. Bordwell, 2007). Notions of cohesion as developed within linguistics have consequently been used to characterize communicative properties in a variety of media, including film (van Leeuwen, 1991; Janney, 2010), text-image relations in printed texts (Royce, 1998), comics and graphic novels (Stainbrook, 2016), and others. Each of these show, in rather different ways, how the media addressed exhibit phenomena corresponding with the distinct types of linguistic cohesion set out by Halliday and Hasan (1976). In a similar vein, van Leeuwen (2005) introduces some additional

* Corresponding author.
E-mail address: tseng@uni-bremen.de (Chiao-I Tseng).

https://doi.org/10.1016/j.dcm.2021.100544
2211-6958/

'multimodal cohesion' categories, including rhythm and composition. Although descriptively interesting, many questions remain concerning the functional role of such cohesive-like devices in comprehension and production, particularly when moving to consider multimodal communicative forms.

Distinguishing the potential contributions of different types of cohesion becomes increasingly important here because each type involves rather different mechanisms and appears to perform distinct discourse functions. Martin's (1992) explicit placement of cohesion at a 'higher' level of descriptive abstraction, conceptually distinct from the particular linguistic forms by which cohesion occurs, establishes an excellent position from which to generalize cohesion beyond verbal semantics, while still maintaining links to operationalisable traces in form. In this paper, we present work extending this line of investigation further for the audiovisual medium of film.

Film is a particularly appropriate target for exploratory multimodal cohesive analysis in several respects. First, films regularly combine spoken and written language, sound (musical, natural and designed), movements, and other visually-carried information such as points-of-view, gestures, facial expressions, proximity and so on (cf., e.g., Bordwell, 2007). Since all of these properties are actively deployed in a deliberately integrative fashion, they demand consideration of a significantly broadened notion of text as multisemiotic communicative device. Second, films are, despite their semiotic complexity, still primarily *linear* expressive forms in that they unfold strictly in time. This provides a solid basis for a close contrastive investigation of similarities and differences with verbal language, which is similarly linear. And third, there is a growing body of work probing commonalities in the cognitive and neural processes of discourse comprehension exhibited in response to language and to film. Here, earlier models based on segmenting the meanings gained from text into events during comprehension (Zwaan and Radvansky, 1998) are finding broader application for the comprehension of narrative in several media, including film (Zacks and Magliano, 2011; Zacks et al., 2007; Kurby and Zacks, 2008; Radvansky and Zacks, 2017).

Such experimental studies show compelling empirical support that text interpreters, in both verbal language and film, closely track (changes in) locations, times, participants and causal relationships to achieve discourse comprehension. When core features of the situation change, a 'current event model' has to be updated into a new model and readers, audiences, and perceivers in general experience this as an event boundary, with corresponding consequences for memory and processing (Zacks et al., 2009). Such features also relate to elements pursued in cohesion analysis. We propose, therefore, that conducting systematic cohesion-oriented analysis may further contribute to, and interact with, studies of this kind. In essence, this prepares the ground for a range of empirical investigations of film comprehension guided by cohesion theory. Conversely, broadening the descriptive and empirical bases of accounts of cohesion should also deepen our understanding of the workings of such discourse mechanisms generally, including those mechanisms engaged during the comprehension of verbal language. Multimodal cohesive analysis may then allow us to triangulate discourse phenomena across several expressive forms, with beneficial consequences for our understandings of each.

In order for such investigations to proceed, however, it is crucial that evidence be found that abstract cohesive analyses and actual processes of discourse comprehension can be systematically related. In other words, it should be possible to show empirically measurable differences in discourse comprehension that align with variations in accompanying patterns of cohesion. If no such connections can be determined, then there would be few grounds for employing cohesion analysis as a means for characterising filmic discourse organisation. Our focus in this paper is therefore precisely that of probing the effect of variations in cohesive organisation empirically. More specifically, we address the empirical consequences of variations in cohesive patterns in specifically modified film sequences. The modifications we make are motivated entirely by an audiovisual cohesion analysis in order to provide materials that differ according to regular differences in cohesive organisation.

The paper is structured as follows. First, we introduce the approach to filmic cohesion that we employ, explaining how this moves beyond the typical usage of the term *cohesion* in film analysis by adapting the discourse-functional perspective of Martin mentioned above. Second, we explain the kinds of modifications we made in selected film segments so as to provide experimental materials, showing how the modifications are systematically described in terms of cohesion. Third, we present the results of questionnaire and eye-tracking experiments, and then go on to discuss implications we draw for the discourse organisation of film and its consequences for processing. Finally, we conclude with further more general comments about the kinds of uses that might be made of cohesive analysis in the future, building on the results presented here.

2. Multimodal cohesion as a component of filmic discourse

It is well-known from film theory that the audiovisual medium of film makes particular use of repetitions, re-occurrences and similarities in forms, both in order to help guide the audience's construction of coherent interpretations of the materials being processed and in order to encourage emotional and aesthetic engagement (Bordwell, 2006; Bordwell, 2007). Such devices range from visual parallelism to support continuity when shifting between shots, over particular patterns or sequences of repeated framings, to musical motifs indicating the recurrence of certain characters or events. The term *cohesion* as a relatively informal indicator of stylistic repetition finds ready application here and so appears in several discussions of filmic organisation (cf., e.g., Palmer, 1989; Bordwell, 2006). More direct use for the analysis of film of the particular kinds of cohesion articulated for verbal language by Halliday and Hasan (1976) is explored by Janney (2010), who also in Janney (2012) considers several further connections between notions of pragmatics from linguistics and possible mechanisms of filmic discourse.

One of the issues raised by Janney concerns the very different natures of the verbal and visual contributions to filmic discourse. Whereas certain linguistic cohesive relationships, such as for example lexical conjunctive relations, appear to require reference to conceptual schemes of differentiation, Janney notes that visual similarities, contrasts, and repetitions appear more directly and immediately accessible—that is, "[t]here is the possibility ... that cohesion in film discourse is not primarily a conceptual phenomenon at all but rather originally a *perceptual* one" (Janney, 2010, 264; original emphasis). Such concerns reoccur in several discussions of the potential discourse nature of visual materials and demonstrate that a more rigorous semiotic characterisation of the phenomena at issue is essential to avoid confusion. As a case in point, Janney shows several sequences of shots taken as illustrating lexical conjunctive cohesive relations such as 'causality' and 'contrast', and suggests that differences in form (e.g., a gun being pointed followed by a shot of someone wounded, or a view of a smaller person followed by view of a larger person) are immediately accessible perceptually, and it is only on the basis of such information that the actual conjunctive cohesive connections can be derived, even though the images themselves contain no explicit markers that this is what is intended. The lack of explicit markers is a major difference to the situation with verbal language cohesion,

since there the concern is precisely to characterise those textual elements that function cohesively; in film, in contrast, it is unusual for many such relations to be explicitly signalled.

Martin's (1992) development of an additional level of linguistic description concerned specifically with discourse semantics separates out the phenomena formerly grouped under cohesion in an arguably more usable fashion. Conjunctive relations are best aligned with coherence or discourse relations which often, even in verbal language, appear without explicit markers; early applications of conjunctive cohesion to film include van Leeuwen (1991), while functional and formal discourse relations building, in part, on conjunction are developed for film in Bateman (2007) and Wildfeuer (2014). The discourse area corresponding to *referential* cohesion is quite distinct, however, both in performing a different kind of discourse work, i.e., introducing and tracking discourse participants, places and objects within any event sequence, and in relying on explicit markers in the filmic text that can be identified and built upon during analysis. This latter property is particularly valuable when considering operationalisation for empirical studies. Nevertheless, although the general conception of referential cohesion as showing how discourse entities can be introduced and tracked across a text is maintained, film and video offer a far broader range of communicative devices capable of bringing this about.

In contrast, then, to primarily linguistic approaches to audiovisual media that focus on depicted verbal performances and situations within film, TV or video (e.g., Piazza et al., 2011; Bednarek, 2018), the approach here centres on the audiovisual material of film itself as an expressive resource. We are not concerned with verbal interaction in film but with how film techniques themselves operate to guide comprehension and (multimodal) discourse interpretation (cf. Bateman et al., 2017). A cohesive analysis of audiovisual texts in this sense proceeds by picking out how the deployment of image, sound, verbal language, written language, camera movement, framing, colour, and many more operates together to introduce and track event participants, places and objects within any unfolding event sequence. This extension of the notion of cohesion is introduced in detail in Tseng (2013) and forms the foundation for the analysis, manipulations and interpretations that we report on below. Within this framework, cohesive mechanisms in discourse in general, and in film in particular, are seen as strategies for leading viewers in particular directions of interpretation rather than others when attempting to comprehend events in verbal and audiovisual media. As a consequence, we consider analyses of this kind as particularly suited to highlighting the *textually constructed* unity of any audiovisually-depicted event.

We exemplify the process of constructing a multimodal cohesive analysis of film using the beginning of Alfred Hitchcock's *The Birds* (1963); we also employ this segment below in our experimental studies. As with the cohesive model of verbal discourse, analysis proceeds by identifying elements and determining the types of cohesive connections holding between those elements. Such connections are termed cohesive links, or ties, and are classified according to the particular strategies for introducing and tracking discourse entities employed. The classification system developed for audiovisual cohesion by Tseng (2013) is shown in Fig. 1, represented as a *system network*. Networks of this kind are the standard notion used in systemic-functional linguistics to capture the abstract paradigmatic 'choices' available for language users drawn from the meaning potential of their language (Halliday and Matthiessen, 2004) and are applicable to any level of linguistic abstraction.

System networks are read as follows. Sets of contrasting options, called *features*, are organised into individual points of choice, called *systems*. A feature from one point of choice may lead on to further points of choice, creating a network organisation. Single features may also lead to multiple further points of choice (indicated by right-facing braces). Points of alternation may be referred to either by name (e.g., SALIENCE) or by listing the contrasting features (e.g., [presenting/presuming]). Only one feature may be selected at any

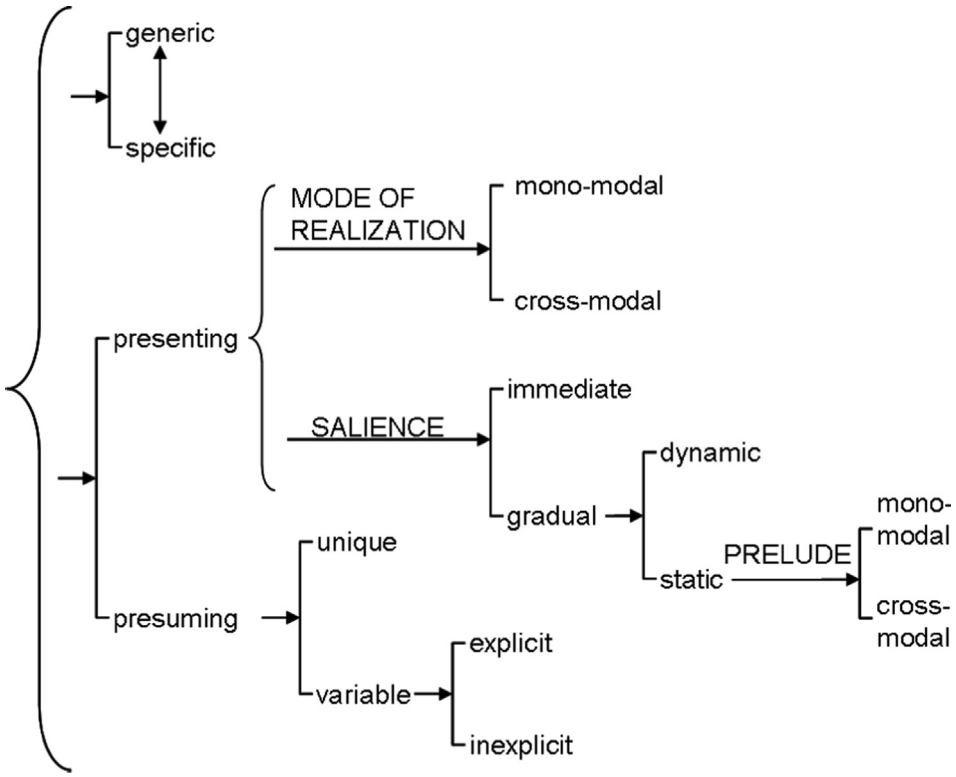


Fig. 1. The filmic identification system network developed in Tseng (2013).

point of choice. In the case of the [generic/specific] alternation in the figure, however, the choice is seen as lying on a continuum rather than being strictly exclusive. When features are not selected, then all points of choice dependent on those features are no longer available. A complete description results when all possible choices compatible with the connectivity of a network have been made. The present network thus shows the conventionalised functional potential for *filmically* cuing identities of event participants, objects and locations as an audiovisually-depicted event unfolds and underlies the particular cohesive strategies discussed in the rest of this paper.

Fig. 2 presents selected stills from the first 40 s of Hitchcock's *The Birds* that we will now subject to a multimodal cohesion analysis following the classification options identified. The film opens by moving the main female character of the film, later identified as 'Melanie Daniels', from the narrative background to the narrative foreground. First, a streetcar is shown passing in image 1, revealing in image 2 a group of people standing waiting to cross a busy intersection. In images 2–6, a female character is successively individuated from the crowd. In image 2 she first stands among the group of waiting people at a distance from the camera, in image 3 she walks left in the image, becoming isolated and thereby visually salient. In image 5 she walks behind a San Francisco tourism poster, emerging finally as a foregrounded character in a medium shot. In images 9 and 11 she is seen noticing the squawking gulls shown in the distance in image 10, before entering a pet shop in image 12. Image 13 then shows her progress within the shop.

Focusing for the purposes of illustration on this female character, we can describe the cohesive devices for presenting and tracking her identity instantiated from the system network of Fig. 1 as follows. In image 2, a female character held in shot is presented for the first time and thus, for this image, it is appropriate to make the choice of the feature [presenting] from the system [presenting/presuming] to capture this. Moreover, she is presented only visually (not simultaneously in written or spoken text), and so the [mono-modal] realization from the system [mono-/cross-modal] is also selected. The foregrounding of her appearance as she moves from the background to the camera's clear focus is gradual and hence this is a realization of the cohesive strategy of [gradual salience] rather than [immediate] in the system SALIENCE. By these means, each re-occurrence of a cohesive element is related to preceding occurrences by specifically labelled cohesive ties showing the identification and tracking strategy involved.

Whereas cohesive ties relate pairs of cohesive elements, sequences of element occurrences and the classified cohesive ties between those occurrences build up *cohesive chains*. Cohesive chains show textual development across larger portions of a text, just as is the case in descriptions of verbal discourse but now extended to include occurrences in any presentation 'modality'. Moreover, in work on verbal texts, it has been observed that such chains and, in particular chain *interactions*, appear to be more revealing of a text's organisation than elements that occur in relative isolation (Hasan, 1984). Interactions between chains occur whenever elements of distinct chains are brought together within the depiction of a single action or event. Thus, although any element in a textual artefact typically enters into a large number of cohesive links with other elements, it is the elements participating in chain interactions that are constructed as being textually 'significant'. This establishes a robust method for selecting from all the cohesive ties potentially available in a text just those collections of ties that are hypothesised to be most likely to play a role in guiding discourse interpretation. That is, a viewer does not need to attend to 'everything' that is audio-visually on offer, but rather will be guided to attend to those elements that contribute to interacting chains.

Performing the same process of identification of elements and classification of the cohesive strategies used to track those elements within the example extract consequently reveals five promi-

Fig. 2. Selected frames from the opening sequence in The *Birds* (1963).

nent narrative elements of character, object, and setting: the San Francisco *street*, the female *main character*, *people* on the street, *birds* and a *petshop*. These identity chains and their constitutive cohesive ties are shown running vertically in Fig. 3, with the rows indicating temporal co-occurrence of elements. Individual ties are shown as arrows where the occurrences of elements 'point back' anaphorically to previous occurrences; although omitted in the figure, each anaphoric tie is assigned a specific label from the classification network of Fig. 1 as well. All of the elements identified exhibit interactions with elements of other chains: for example, the main character 'looks at' the birds, 'goes into' the petshop, 'walks in' the street, and so on. The many other narrative elements that might have been included in a cohesion analysis simply by virtue of their presence in frame fall away at this point, precisely because they do not participate in chain interactions. Further details and examples of the method of analysis can be found in Tseng (2013).

The identified chains can be related straightforwardly to the construction of 'events' introduced above. Within the first portion of the analysed segment, for example, the cohesive chains of 'street' co-pattern with the identity chains of the female 'main character', 'people' and 'birds' to construct a coherent event that might be glossed in natural language as follows:

> The female [main character] walks with other [people] on some [street] of San Francisco, seeing some [birds] squawking and flying in the distance.

Similarly, the cohesive chains of the latter portion of the segment allow the construction of an event:

> The [female character] goes into a [petshop] with some chirping [birds].

We predict that viewers are likely to see such 'abstract' event configurations arising naturally out of the texture of the audiovisual material they engage with. In this sense, therefore, the deployment of the material possibilities of film itself serves a central role in guiding a film's reception. This means that we would predict that variations in cohesive chains and their interactions should lead to corresponding alterations in an audience's understanding. We now probe this experimentally.

3. Employing empirical methods to test multimodal cohesion in event comprehension

We have argued that cohesion analyses of an appropriate kind can show how the audiovisual elements presented and maintained in a segment of film might guide interpreters to construct particular 'events' on the basis of the cohesive cues given. While plausible on abstract grounds, it remains to be seen whether we can find empirical support for such a close association of cohesive patterning and interpretation. To pursue this, we apply the methodology of selecting film segments and systematically modifying those segments so that they exhibit different patterns of cohesion. The segments, original and modified, are then shown to different groups of

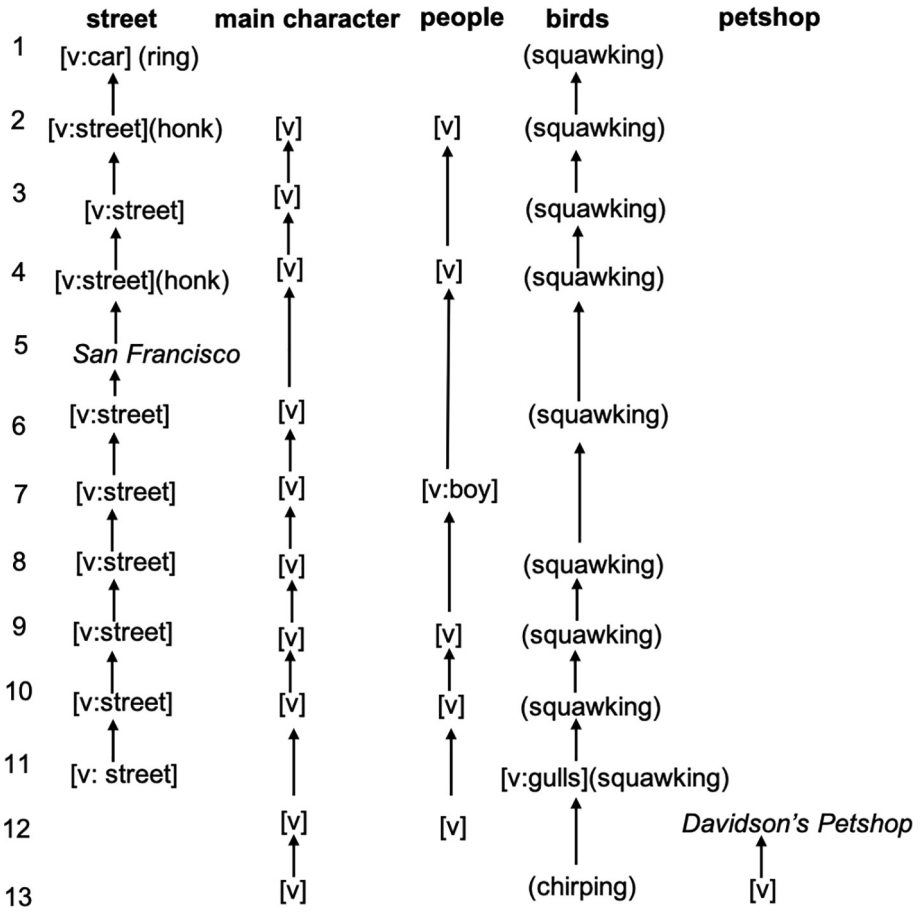


Fig. 3. The cohesive chains of the beginning sequence in *The Birds*. The numbers refer to the images in Fig. 2. [v] refers to an identity realised in visual mode, italic to realisations in written text, and to aud.io realisations.

participants, and differences in their engagement with the materials are measured. In the work reported in this paper, this measurement was performed in two ways: first, by presenting questionnaires concerning the events that the experiment participants took themselves as having seen, and second, by conducting eye-tracking experiments to determine whether different gaze patterns, and hence allocations of attention, follow in the unmodified and modified conditions.

Two studies testing these hypothesized functions of cohesive cues were performed. The first study used the sequence from *The Birds* analysed above; the second applied the same method with respect to the opening of the *Monty Python* TV series sketch entitled 'Dirty Hungarian Phrasebook'; for the purposes of our empirical study we used the filmed version of this sketch from *And Now for Something Completely Different* (1971, directed by Ian MacNaughton). In both cases, modified versions of the segments were created and then questionnaires and eye-tracking experiments were carried out with participants who had watched either the unmodified version or the modified version. The modifications to the segments were made in a way that rendered them undetectable to any viewer who did not know the original segment (and often even then). The contrasts and the results they led to are set out below.

3.1. Manipulation of the opening segment of The Birds

For the manipulation of the opening segment of *The Birds* we focused on the event construction of the latter portion where the female character enters the pet shop. As shown in Fig. 3, the original segment includes cohesive cues that explicitly identify the kind of shop that is being entered. We therefore subtly removed the verbal and audio cues which indicated its specific identity, namely, blurring the written signs identifying it as a pet shop and replacing the bird chirping sounds within the shop with generic soft background music. The segment inside the pet shop was therefore visually identical across the two conditions and the modification of the scenes is not generally perceptible to casual viewers, who usually fail to notice that the scenes differ at all. Our hypothesis, however, was that disrupting the cohesive connections in this way should nevertheless have consequences for interpretation and so variation in gaze behaviour was predicted to occur even for the visually unchanged portion of the segment. Here it is important to note that it is generally not straightforward to trigger gaze behaviour in film-viewers that deviates from that predicted on the basis of visually-present action cues (cf. Loschky et al., 2015a; Kluss et al., 2016) – achieving measurable differences for visually unchanged portions of film is thus challenging; we return to this below.

Fig. 4 shows the audiovisual cohesive chain analysis of the modified segment. The removal of the pet shop signs (e.g., the written text of 'Davidson's Pet Shop', shown on the right of the figure) results in the change of the chain indicating the setting from a specific named pet shop to some generic shop. In addition, the removal of bird chirping sounds inside the shop breaks the continuing cohesive chain of 'birds', that is, the birds were not tracked in the audio mode after the main character enters the shop. In terms of multimodal cohesion and the classification system of Fig. 1, therefore, the modification undertaken at the discourse level was a change in presentation strategy for the shop setting from [specific] to [generic].

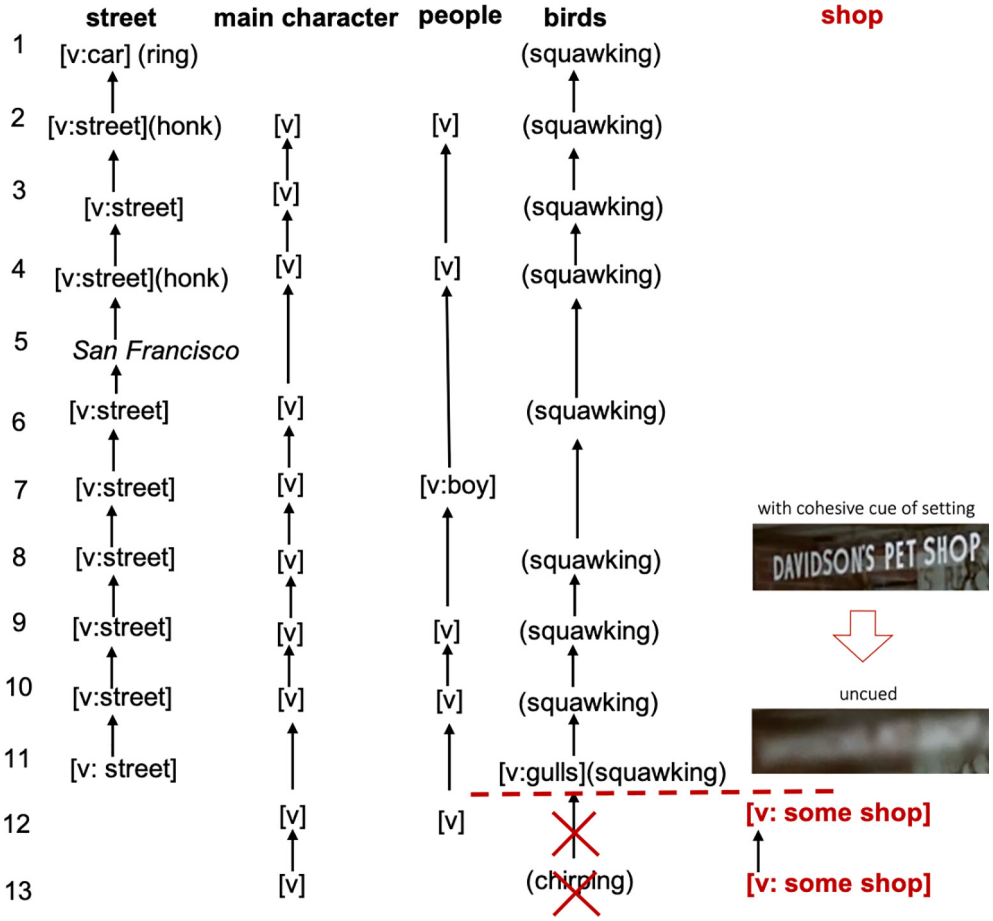


Fig. 4. Changes in the chain patterns of the manipulated version of the opening segment of *The Birds*.

3.2. Manipulation of the opening segment of the 'Dirty Hungarian Phrasebook' sketch

The film material for the second study is displayed in Fig. 5 in a similar fashion to that used for the previous example. This beginning segment of the sketch shows the interaction of a 'Hungarian man' (John Cleese) and a shopkeeper (Terry Jones). Performing a cohesion analysis of this segment following the principles given above establishes the overall set of chains shown in Fig. 6a. As before, the chain pattern explicitly presents and tracks the narratively significant people, places and things in the segment. In this case, the two main characters, 'The Hungarian man' and the 'seller', the specific setting, i.e., a 'tobacconist', and the object 'phrasebook' seen in the Hungarian man's hand throughout the sequence are unproblematic. Several other shorter chains track the themes mentioned in accompanying verbal texts, including a voice-over introduction concerning the time and location, 'London 1971', as well as the objects 'hovercraft' and 'eels' mentioned by the characters during the sketch. The generic customer visible in image 2 of Fig. 5 is not included in the chain figure because of his non-dominant presence: he is not seen frontally and is only briefly, non-repetitively presented. As a consequence, there is almost no cohesive chain interaction and he serves more to contextualise the overall buying-selling setting rather than being a character. We shall see below that this analysis is also confirmed by the eye-tracking data, which shows that the customer indeed barely attracted any attention (cf. Fig. 12).

A manipulated version of this segment was created by again deleting elements that made explicit the specific identify of the shop setting. To achieve this, we removed those parts of the verbal texts in the voice-over referring to tobacconists and also cut out the sequence between image 5 and image 7 as well as the last part of the dialogue shown in image 8, in which the seller points to the background shelf and takes a pack of cigarettes and matches, mentioning tobacconist-relevant objects, such as cigarettes and matches. Similarly to the bird cages in *The Birds* manipulation, however, we did not remove the many packs of cigarettes shown in the background as we assumed that, in the absence of further guiding cohesive chain interactions, these would also not be seen as prominent cues. The elements extracted are also shown in Fig. 5, while the cohesive chains corresponding to the manipulated version are shown in Fig. 6b. Compared with the chain patterns of the original version, we can see that the original setting chain 'tobacconist' has become a generic shop chain because the specific identity of the shop is no longer explicitly cued—this is consequently parallel to the manipulation applied to *The Birds*.

In both cases, the original and manipulated versions in the two studies were then used to test if the respective versions differed in terms of their take-up by viewers. In particular, we hypothesized that there should be consequences for the event settings constructed in the segments as these are where the cohesion analyses differ.

3.3. Comprehension test

As indicated above, the association hypothesized between cohesive chain patterning and event construction as a component of interpreting the audiovisual material should, if accurate, lead to our manipulated segments having particular interpretative consequences for their viewers. These consequences would not necessarily be evident to viewers but would, if present, affect their online interpretations. As a first step, therefore, we assessed whether the manipulations indeed had effects of the kinds we predicted. This led to the specific hypothesis:

- Viewers of the manipulated versions in both studies will be less certain about the specific identities of the shop, even though the relevant visual elements inside the shop are still readily accessible on screen.

This was investigated by having participants answer the following questions immediately following their viewing of the respective film segments:

- *The Birds*: "What does the character walk into"?
- *Monty Python*: "Where does the dialogue take place"?

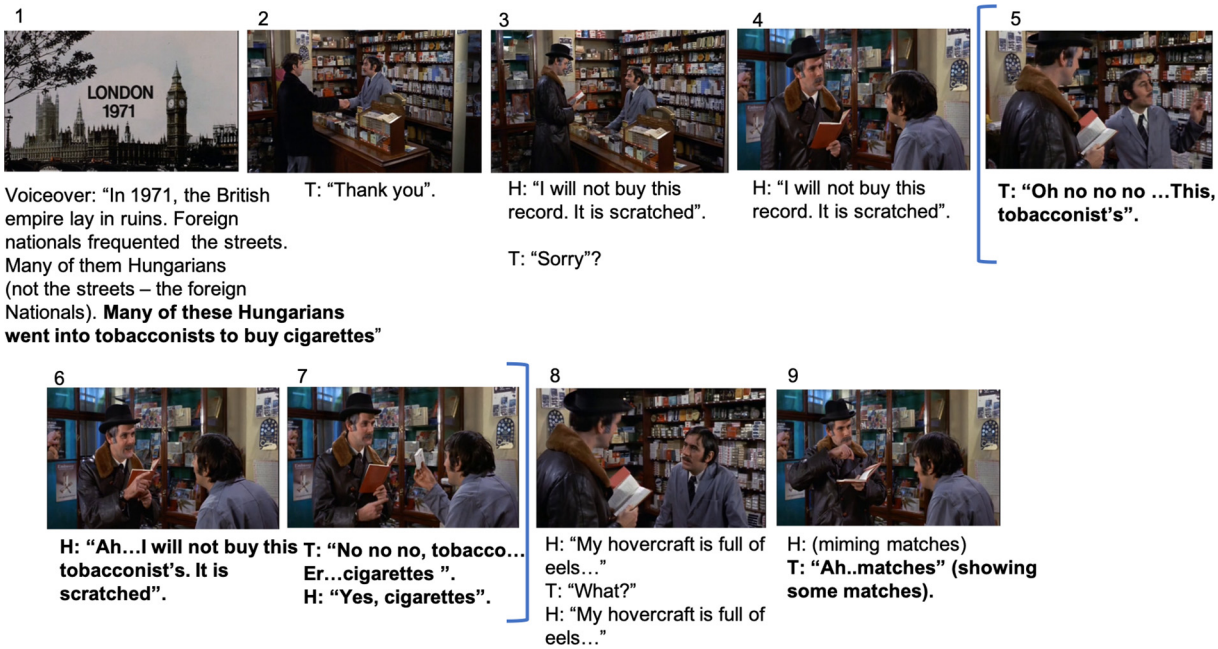
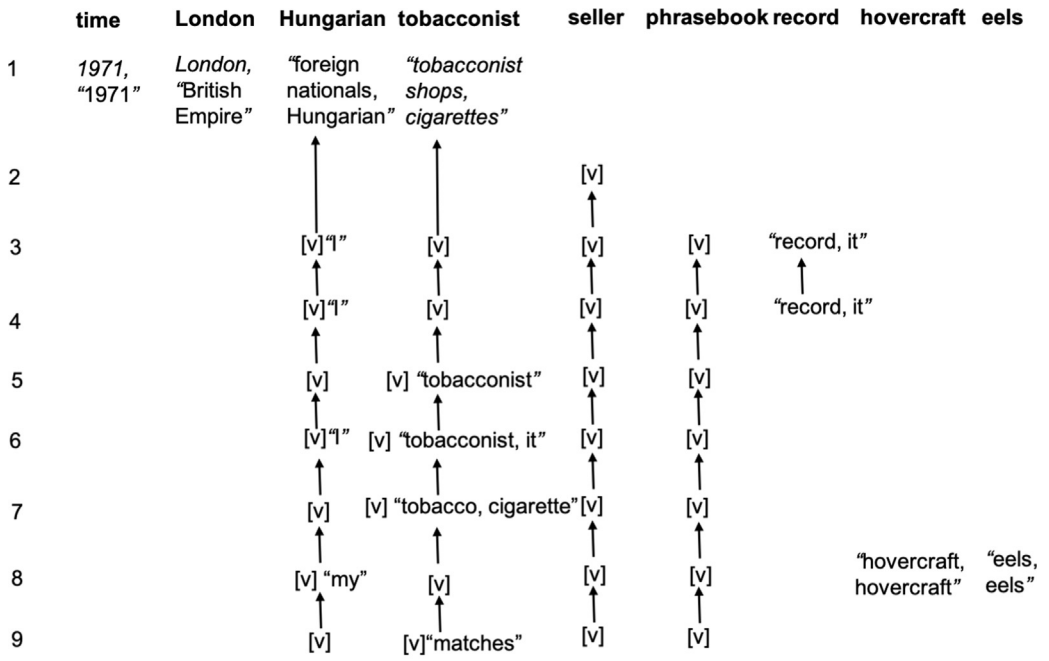
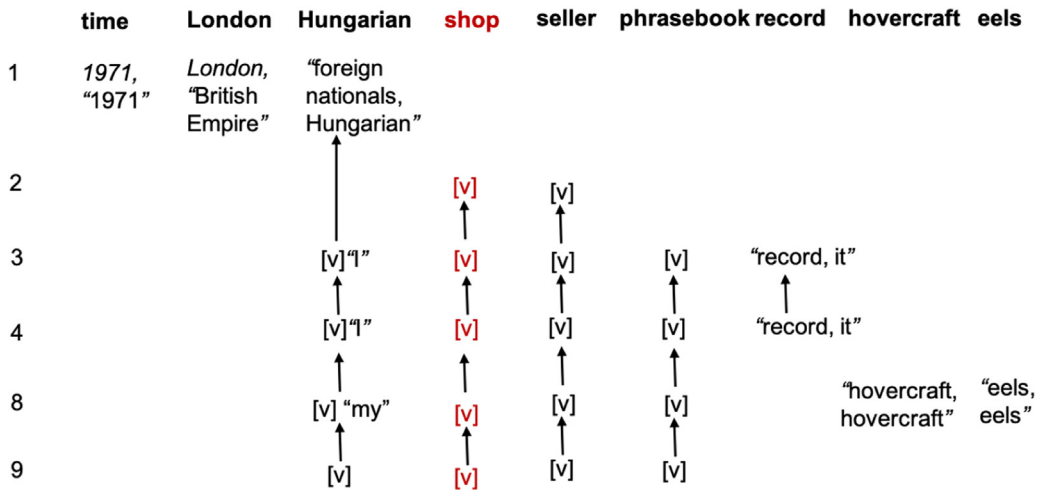


Fig. 5. The original and manipulated version of the beginning of the *Monty Python* sketch "Dirty Hungarian Phrasebook" (taken from the 1971 film *And Now for Something Completely Different*, directed by Ian MacNaughton). In order to remove cohesive cues of the specific setting, the bold verbal texts and the sequence shown from image 4 to image 6 are cut in the manipulated version. H: Hungarian, T: Tobacconist.



(a) Original version.



(b) Manipulated version with the 'tobacconist' chain becoming a generic 'shop' chain.

Fig. 6. Cohesive chains of the original and modified versions for the *Monty Python* sequence.

Two sets of comprehension tests were conducted. The first set was conducted at the University of Bremen, and the participants ($n = 45$) were undergraduate students who had not seen the segments before the experiment. The participants were divided into two groups. Group 1 ($n = 23$) watched the original versions (i.e., the cohesively 'cued' versions) of the two sequences, while Group 2 ($n = 22$, 'uncued') viewed the manipulated versions with cohesive cues removed. A replication of the comprehension tests was conducted at the University of Potsdam in the context of the eye movement experiments described in Section 3 below. The questionnaire results obtained from $n = 74$ and $n = 34$ participants for *The Birds* and *Monty Python* respectively, will be reported in the present section as well. Fisher's exact test was used to evaluate statistical significance of dependencies between the cohesion status (cued *vs.* uncued) and viewers' interpretations of the location (correct *vs.* incorrect), with $p < 0.05$ considered significant.

Table 1a presents the questionnaire results for *The Birds* separately for the Bremen and Potsdam samples. In the Bremen sample, all 23 participants in Group 1 who watched the cued version were aware of the specific identity of the pet shop, while only 16 participants from Group 2, who watched the uncued version, answered the question correctly. The 6 participants who were not certain about the location gave answers varying from a generic store to a boutique or a company. Fisher's exact test showed a significant association between the "variables" cued/uncued and correct/incorrect ($p = 0.0092$). Thus, although it is certainly the case that viewers of the uncued version might be able to guess the kind of

Table 1
Questionnaire results for the two comprehension studies.

(a) *The Birds* study: Number of participants with correct or incorrect answer to the question in group 1 (cued version) and group 2 (uncued version)

Bremen sample

	Cued	Uncued	*Total*
Correct	23	16	39
Incorrect	0	6	6
Total	23	22	**45**

Potsdam sample

	Cued	Uncued	*Total*
Correct	31	23	54
Incorrect	6	14	20
Total	37	37	**74**

(b) *Monty Python* study: number of participants with correct or incorrect answer to the question in group 1 (cued version) and group 2 (uncued version)

Bremen sample

	Cued	Uncued	*Total*
Correct	23	3	26
Incorrect	0	19	19
Total	23	22	**45**

Potsdam sample

	Cued	Uncued	Total
Correct	16	2	18
Incorrect	1	15	16
Total	17	17	**34**

shop involved correctly based on the remaining available information such as the cages in the background, the question interrogated here is whether the manipulation makes a difference. The results demonstrate that the cued and uncued versions indeed differ significantly in comprehension. Questionnaire results from Potsdam confirmed these results. The earlier results obtained in Bremen justified the use of a one-sided test, giving a significant result ($p = 0.033$). Inspection of the qualitative results suggests that in the uncued version, the shop identity was sometimes interpreted to be more generic, e.g., a department store.

Table 1b presents the results of the *Monty Python* study. In the Bremen sample, all 23 participants in Group 1 who watched the cued version were aware of the specific identity of the tobacconist, while only 3 participants from Group 2, who watched the uncued version, answered the question correctly. In Group 2, 19 participants were not certain about the location and most of them answered "a shop" or "pharmacy". Fisher's exact test again indicated a statistically significant association between the two variables ($p < 0.001$). The Potsdam results confirmed these results further. Removal of the voice-over phrase and other cues to the tobacconist resulted in considerable confusion about the identity of the shop ($p < 0.001$). Inspection of the qualitative results suggests that the large effect is partly due to the fact that British tobacconists in the seventies look like pharmacies to the German viewer from the late 2010s.

Drawing on the two questionnaire studies, the hypothesis that the manipulation of verbal and audio cohesive cues would affect the viewers' comprehension of the depicted event's location is supported. In both cases, viewers of the manipulated versions were less certain about the specific identity of the setting inside the shop than viewers of the original versions. Confusion about the shop identity was numerically stronger with *Monty Python* than with *The Birds*, potentially due to a confounding similarity with modern-day pharmacies in Germany. In summary, the questionnaire data of the two studies suggests that the manipulated cues were crucial in specifically indicating the event settings.

3.4. Eye-tracking experiments

In the second set of studies, we performed eye-tracking experiments for the two pairs of original and unmodified sequences described above in order to further refine our understanding of the consequences of the manipulations.

Attention and gaze position as measured by eye-tracking are generally thought to be driven bottom-up by the visual and auditory stimulus as well as top-down by cognitive and interpretative processes and expectations (Itti et al., 1998; Wolfe, 1994; Torralba et al., 2006; Corbetta and Shulman, 2002; Schütt et al., 2019). Earlier studies of eye-tracking on film have shown, however, that the dynamic nature of the medium provides sufficiently strong attentional cues that top-down processing, such as that which might be predicted according to discourse processing mechanisms, is largely overruled. In addition, cues employed by directors and cutters concerning continuity and similar techniques of film-making drive viewers gaze behaviour still further, leaving rather limited room for top-down processes and inter-individual variability (Smith, 2012). The distribution of attention in film being largely determined by bottom-up processing has been termed the "tyranny of film" (Loschky et al., 2015a; Loschky et al., 2015b). In contrast, effects of cohesion, as discourse phenomena that are brought to material during interpretation, are top-down effects. If, therefore, we were to find the predicted top-down effects of our manipulation of cohesive chains, this could be regarded as strong evidence for an effect of cohesion on comprehension – strong enough to guide attention against the dominant visual cues.

3.4.1. Experiment 1: The Birds

As explained above, for the first experiment we edited the beginning of *The Birds* to remove visual cohesive cues to the pet shop from the establishing shot outside the shop, whereas the subsequent scene inside the pet shop remained visually identical across the two conditions: original (cued) and modified (uncued). We then collected eye-tracking data of a total of 114 participants to measure the spatio-temporal distribution of attention.

Two groups of participants watched either the original version or the manipulated version. We had the specific hypotheses that (a) the deletion of visual cohesive cues in the manipulated version would lead to less attention directed towards their respective locations in the establishing part, and (b) more orienting behavior would be required by participants when the specificity of the shop had not been established by cohesive cues. As a consequence, a broader (less focused) distribution of attention was expected in the scene inside the pet shop.

Methods. Two samples of participants took part in the experiment. The first sample consisted of 34 participants (27 female, median age 22 years, range 19–40 years), who also watched some other movie clips. The second sample consisted of 80 participants, who were recruited to test a hypothesis derived from the first sample's results, and who only watched clips of *The Birds*. All participants were drawn from the University of Potsdam subject pool and received financial compensation or course credits. Movie type (cued vs. uncued) was a between-subjects variable. Gaze position was sampled binocularly at 1000 Hz using an EyeLink 1000 eye tracker (SR Research), following a 9-point calibration, which ensured that gaze position was measured with an error of less than 0.5 degrees of visual angle. The clips were presented on an iMac at a resolution of 5120×2880 pixels. Stimulus presentation was controlled using the Experiment Builder software. In addition to the eye tracking measurements, participants also answered a questionnaire on the location of the scene as reported above. Due to a communication error, questionnaire data was missing for 40 of the 80 participants of the second sample.

Results. Generally, eye-tracking revealed differences in participants' gaze behavior between the manipulated and original versions of *The Birds*. First, in the establishing parts of the scene, manipulated scene content (i.e., the blurred version of the pet shop signs) received less attention than its original counterpart. This is indicative of the strong attention-focusing function of text, and might be considered a manipulation check: Viewers of the original version indeed paid attention to the signs establishing the shop's identity, as illustrated in Fig. 7.

During the periods when the signs were visible, about 18% of the fixations in the cued condition and only 1% of the fixations in the uncued condition were on the regions containing the signs. This group difference was highly significant, as shown by a Welch t-test, $t(58.29) = 8.61, p < 0.001$. 83% of cued participants attended the signs at least once, compared to 15% in the uncued condition. Fig. 8 shows that the locations of the blurred signs were hardly ever attended.

Apparently attending the signs indeed served a cohesive function because, second, the distribution of gaze locations on the *identical* scene inside the pet shop differed between groups, showing that the missing explicit verbal cues establishing the setting affected viewers' orienting behavior in the later scene. Fig. 9 illustrates the distribution of attention at one point in time, at which the main character turns her head while walking up the stairs, thereby signaling a shift of her attention. Although most fixations are on either the character's face or the main bird cage, the distribution seems more focused on the character and cage in the cued condition and more spread out throughout the store in the uncued condition.

To formally investigate the effect, we examined the unfolding of the spread of the two-dimensional distribution of gaze locations over time. We assume that more spread is correlated with more exploratory behavior and searching, whereas less spread is associated with focused attention. As a measure of spread in two dimensions we computed the square root of the determinant of the covariance matrix, which can be considered an extension of the notion of standard deviation to higher dimensions (Paindaveine, 2008). Fig. 10 shows how the spread of the distributions developed over time in the shop. In both conditions, the spread was initially increasing, and dropped towards the end of the scene. However, the drop occurred earlier in the cued condition, and appears to coincide with the turning of the head of the main character. We speculate that this is because the main character's turning her head served as an attentional cue to the viewer—a social signal to follow

Fig. 7. Distribution of fixations in the shot establishing the pet shop identity. Whereas viewers often fixated the 'Davidson's pet shop' text and 'pets/birds/tropical fish' signs in the original version (*left*), they hardly ever attended the blurred signs in the uncued version (*right*).

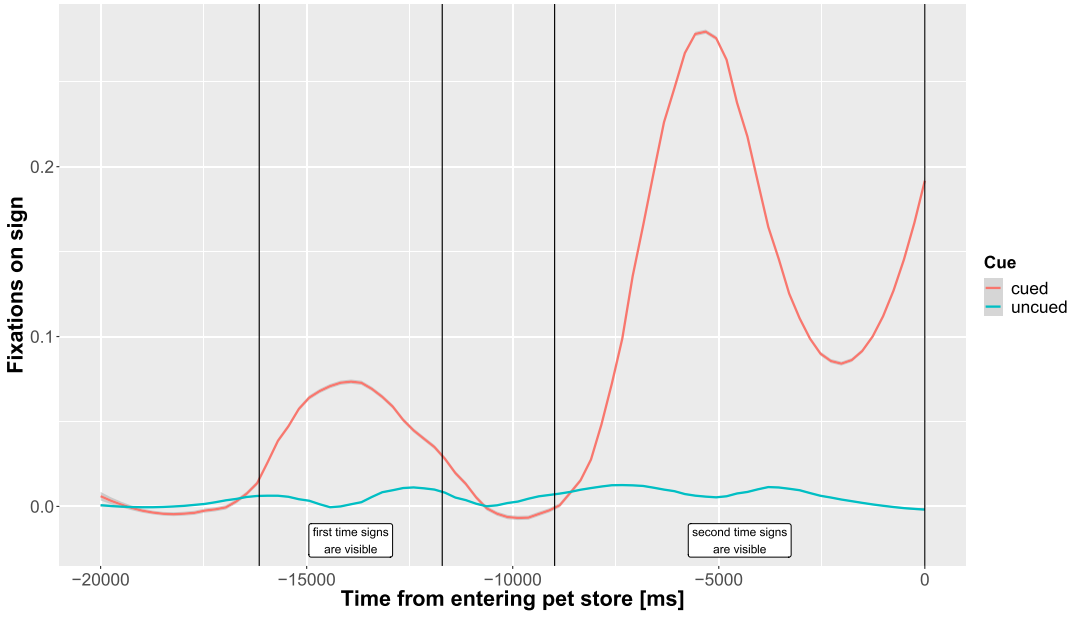


Fig. 8. Fraction of fixations falling on the pet shop signs over time, relative to the main character entering the shop. Clearly, the regions of the signs were (only) attended in the cued version.

Fig. 9. Fixation distributions inside the shop when the main character turns her head, where a large cage of birds is visible, and where later a shop assistant will appear. In the uncued version (right), attention is distributed more widely, so that the main character's gaze cue tends to be missed more often.

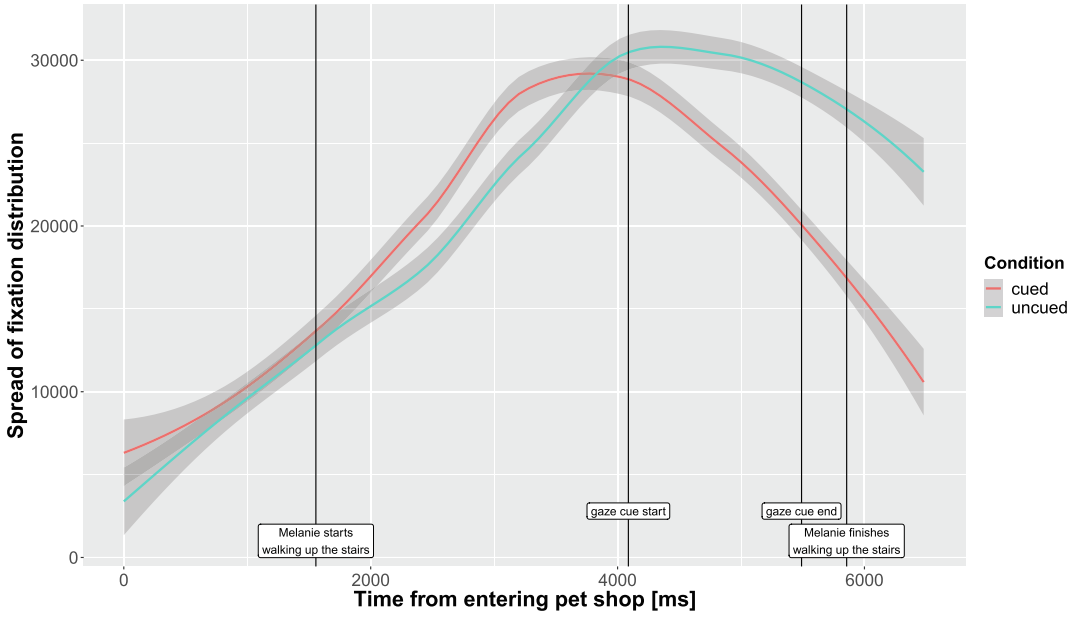


Fig. 10. Spread of the fixation distributions inside the shop over time.

her attention, similar to a gaze cue—and that this cue was attended more often by the viewers in the cued condition, whereas viewers in the uncued condition continued to explore the shop. Summed over the whole scene, the spread of the distribution was clearly larger in the uncued condition, although initially the distributions developed more or less in parallel.

An interesting question is raised concerning where the main character's head-turning cue leads viewers' attention. It is well-known from eye-tracking studies generally that observing someone looking in a particular direction is likely to trigger gazes in that direction also (Friesen and Kingstone, 1998); this can be utilised in film, therefore, as an explicit directional cue for audiences. Fig. 11 shows the number of gaze fixations falling on the cages that become visible as the main character is walking up the stairs. Importantly, attention on the cages was more synchronized in the cued condition, and seemed to be at least partly triggered by the head turning. In contrast, attention to the cages started earlier and was more variable in the uncued condition. Both are indicative of exploratory behavior.

Considering the results of the comprehension questionnaires showing participants less likely to identify the setting location in the manipulated version although visual elements were still dominantly seen on the screen, the eye-tracking data suggest that viewers of the original version indeed used the pet shop signs to establish the identity of the shop, whereas viewers of the manipulated version were later actively searching for information inside the shop to reduce their uncertainty about the setting. To summarize, in this case, verbal and audio modalities functioned significantly to direct viewers' attention and affected their narrative comprehension process.

3.4.2. Experiment 2: Dirty Hungarian Phrasebook

In the second video, the Monty Python sketch 'Dirty Hungarian Phrasebook', the tobacconist setting is established by the voice-over phrase "Many of these Hungarians went into tobacconists to buy cigarettes", which we had erased from the manipulated version.

Methods. The same 34 participants as in the first sample of Experiment 1 took part in the experiment. In a between-subjects design, we presented the original and the manipulated versions to 17 viewers each. All other details were as in Experiment 1.

Results. For the analysis of the eye-tracking data, we focus on the initial scene inside the shop, where an extra is seen buying a pack of cigarettes before the "Hungarian" (John Cleese) enters the shop (image 2 in Fig. 5). This scene serves to visually establish the tobacconist setting, which in the original version was already

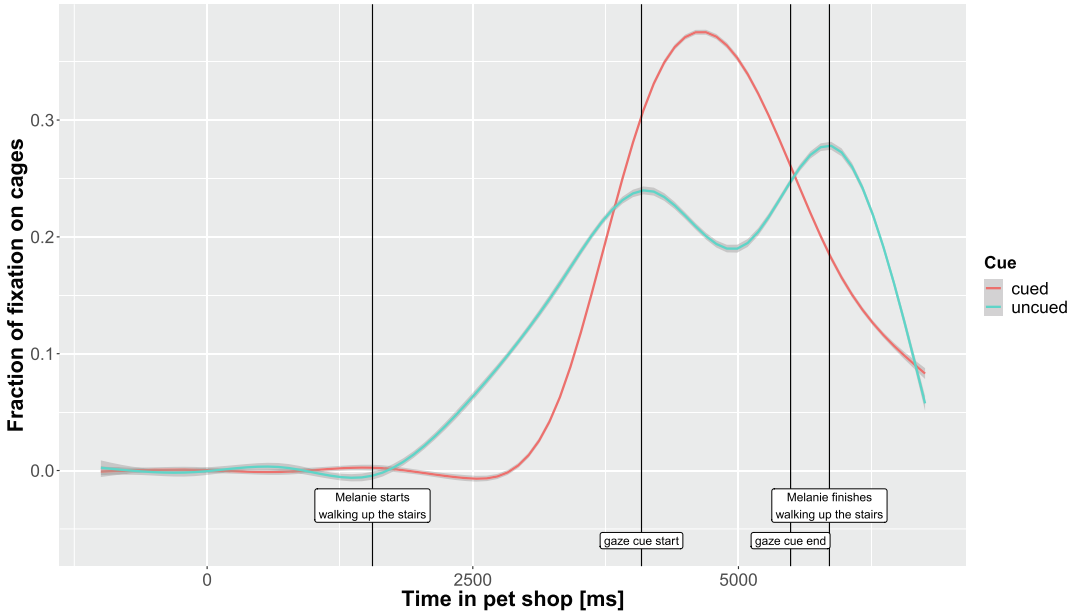


Fig. 11. Fixations on bird cages inside the shop over time.

primed by the voice-over phrase. This scene is ideal for measuring deviations in gaze, because the gist of a scene is usually established in the first few fixations.

Fig. 12 shows that viewers of the manipulated version seemed to be less certain about the specific identity of the setting inside the shop, as indicated by the larger number of fixations on elements such as the cigarette packages on the shelves in the background, and the generally larger spread of the distribution. In contrast, viewers of the original version almost exclusively focused their attention on the tobacconist.

To formally evaluate this effect, we computed the measure of spread of the two-dimensional distribution and calculated bootstrapped confidence intervals based on $N = 20$ replications. The formal evaluation shows that the spread of the fixation distribution was indeed smaller in the original than in the manipulated version ($M = 6265$ vs. $M = 11149$; bootstrapped confidence intervals did not overlap), suggesting that viewers of the manipulated version were actively searching for information in the establishing shot. In contrast, viewers of the original version concentrated on the actors and actions, i.e., handing over the packet of cigarettes. Again, the result suggests that cues to cohesion as employed by the film makers tend to be picked up by viewers. When cohesive cues are missing, viewers need to divide their attention between following the dialogue and finding out where it takes place.

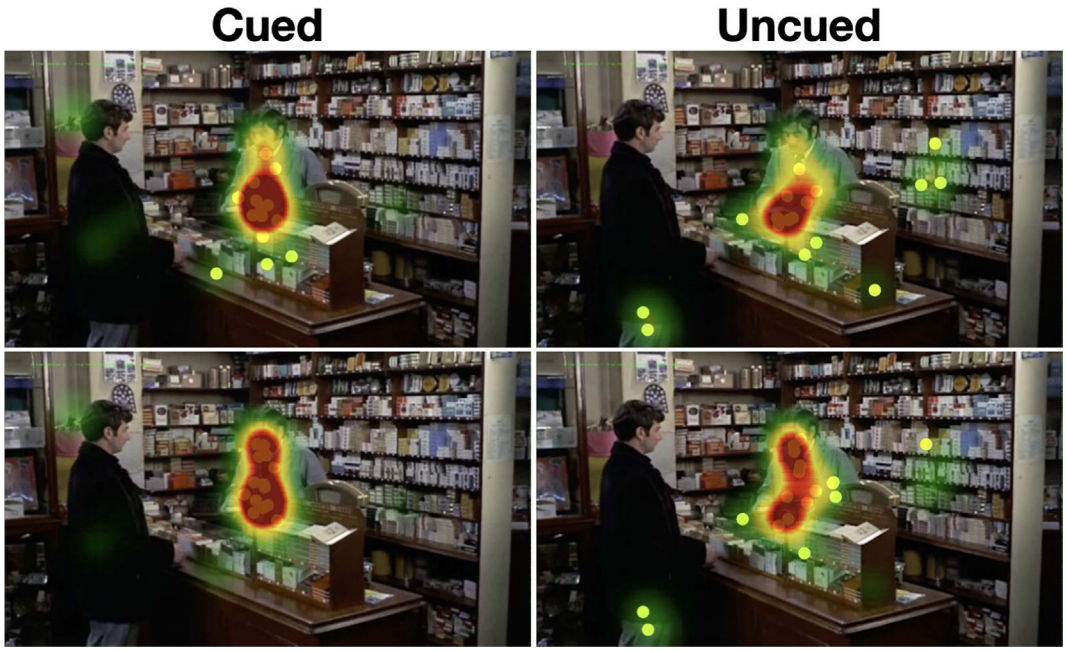


Fig. 12. Fixation distribution at two time points (*top, bottom*) shortly after a customer entered the tobacconist's shop. The tobacconist attracts most of the attention, but less so in the manipulated version (*right*), in which more exploratory behavior was observed.

4. Discussion

In this exploratory study we have presented empirical evidence from two studies that manipulating visual and auditory cohesive cues to the identity of a setting leads to uncertainty about the setting, which triggers active search behaviour in order to reduce that uncertainty. Although rather different physical manipulations were made in each case—one involving written language and sound-effect changes, the other involving spoken language—very similar uncertainties and changes in gaze behaviour resulted. These similarities align well with the single change made in the cohesive strategies employed: i.e., replacing the [specific] identity strategy with a [generic] strategy. Indeed, the stimulus manipulations were theoretically derived from the cohesion framework, and the empirical results were compatible with its predictions. Nevertheless, it is clearly not yet possible to state unequivocally that it is the change in cohesion that brings about the changes in interpretation. An alternative possibility is that participants with less information tend to know less and tend to search more for additional clues. As a consequence, it is still possible that our results might result from a general lack of knowledge rather than from a weakened cohesive chain. Ultimately answering this question will need additional experiments capable of singling out specifically cohesive effects so that such effects can be distinguished from any variations in interpretation due to non-cohesive effects.

While we cannot yet rule out this alternative, there are several grounds that support our interpretation in terms of cohesion. Although manipulating the audiovisual content of a film at any point might be expected to change the state of knowledge of viewers and so lead to different subsequent behaviour, more detailed consideration of the processes of film perception and interpretation raise challenges for such a view. It is not likely, for example, that a simple lack of knowledge leads to the kind of results that we have reported above because viewers will not, in general, know that they lack knowledge and so cannot engage in search: only very specific kinds of gaps in knowledge appear likely to trigger such activity. Online film viewing is a perceptually expensive task and viewers need to allocate stretched cognitive resources accordingly: they generally do not have capacity free for search if it is not 'required'. The issue then becomes one of establishing when search may be necessary.

If elements in film, or similar audiovisual media, are not attended to by a viewer, then they are not available for building subsequent interpretations. Consequently, their absence will also have no effect (as shown in studies of change blindness and selective attention, e.g.: Simons and Rensink, 2005). The question then arises as to how film-makers can construct material that guides viewers' attention appropriately, i.e., directing attention so that the material necessary for intended subsequent interpretation is available. This is a very different situation to that of, for example, verbal storytelling where it is, in general, not a perceptual limitation that may make certain information available or not for building discourse interpretations, but narrative techniques. With film, perceptual limitations can be engaged directly for narrative and other purposes as well. Film-makers know this and, indeed, regularly manipulate such limitations for discourse purposes.

Our hypothesis is that interacting cohesive chains provide precisely the level of description needed to characterise how attention will be directed during film viewing. Gaps in knowledge will not be considered 'gaps' if they are not relevant for the discourse and discourse relevance is constructed filmically by means of cohesive chains. The opening sequence from *The Birds* provides a suggestive example of this process at work. A well known directorial flourish practised by Alfred Hitchcock is his cameo appearances in his films. In *The Birds*, this appearance occurs within the first 40 s and so fell within the scope of our eye-tracking studies. Consequently, we were also able to analyse if viewers paid particular attention to Hitchcock's presence (image 12 in Fig. 2). The eye-tracking result is shown in Fig. 13, where we see that in both cued and uncued versions, viewers' attention stays mostly on the female character just leaving the shot. Despite the fact that the figure of Hitchcock is visually prominent in the shot and is even moving at some speed, he is not attended to.

This might appear counter-intuitive until we consider again the cohesion analysis in Fig. 3. Here we see that the director does not participate in any interacting cohesive chain (apart from the generic background of 'people') and so would be predicted to be non-salient textually—i.e., not specifically relevant for the discourse. He is accordingly, and certainly as he intended, simply not 'seen'. If elements in a film are not constructed as relevant by making them participate in interacting cohesive chains, then they will not be attended to. The settings of the petshop and the tobacconist in our example studies were, in contrast, directly involved in chain interactions with the principal protagonists and so were certainly made relevant for interpreting likely actions of those protagonists, making their [generic] presentation problematic and perceptible as a 'gap'.

In many respects, therefore, we are proposing cohesion as an appropriate way of *characterising* changes in relevant knowledge. Cohesion analyses should consequently group changes in knowledge into equivalence classes that align with types of changes in behaviour. In the studies reported here, we were only able to consider one very specific kind of cohesive variation. Further empirical studies are clearly needed to explore systematically the extent to which different kinds of cohesive relations correlate with differing patterns of interpretation more broadly.

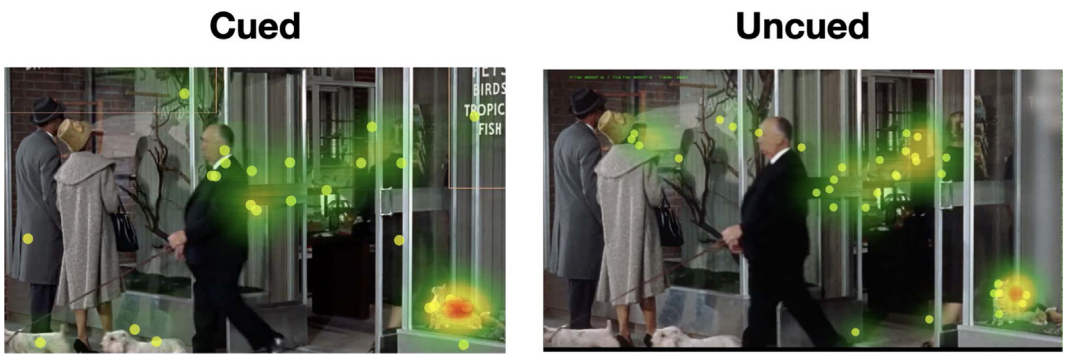


Fig. 13. Hitchcock's cameo appearance was mostly unattended, regardless of cueing condition. Instead, attention tended to remain at the disappearing female character's location in expectation for some new event to appear in the door frame.

5. Conclusions

The two example analyses reported in this paper have given preliminary support to the idea that an appropriately extended multimodal view of cohesion can be applied to explain certain patterns of discourse interpretation in film. Although much remains to be explored, we have shown how cohesion can be given an operationalisable definition that allows film segments to be analysed with sufficient detail to support empirical research. The adoption of such a mixed method approach relies on the important design feature of multimodal cohesion that it is able to systematically establish structures that support the controlled selection and subsequent manipulation of materials for empirical experimentation. Nevertheless, we would say that we are still in the context of discovery in a Popperian sense, and hence exploratory analyses are both justified and necessary.

Finally, several directions for future research suggest themselves.

First, we believe that the empirical framework of multimodal cohesion could be beneficially applied in studies of *event representation*, an issue now receiving a surge of attention across several disciplines (Zacks, 2010; Hafri et al., 2018; Maienborn, 2019). Despite an extensive literature in both verbal and audiovisual media, limited work has been done on the multimodal semantic structure of events *per se* and on the relation between visual, verbal and audio event components. In this respect, multimodal cohesion may be employed to address questions concerning how people perceive and interpret events on the basis of event components represented simultaneously and complementarily in language, image and sound.

Second, the provision of multimodal cohesive chain analyses suggests selection criteria for materials for empirical investigation that may not be directly related perceptually. Finding *comparable* experimental materials is always a challenge as potential confounds need to be minimised. Cohesive chains offer a higher level of abstraction for constructing such contrasts.

And third, multimodal cohesion structures may also be applied across media: this might support empirical studies of the consequences of deploying functionally similar presentational strategies employing elements with quite different affordances. Preliminary investigations involving multimodal cohesion analyses of comics with eye-tracking techniques can be found in Tseng et al. (2018).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to Gary Wong for editing the movie snippets. We thank Jessie Nixon for helpful discussion, Petra Schienmann for data collection, and Dominik Grätz for help in annotating and pre-processing the eye tracking data. JL was supported by BMBF grant 01UG1407B.

References

Bateman, J.A., 2007. Towards a grande paradigmatique of film: Christian Metz reloaded. Semiotica 167 (1/4), 13–64.
Bateman, J.A., Wildfeuer, J., Hiippala, T., 2017. Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction, Mouton de Gruyter, Berlin.
Bednarek, M., 2018. Language and Television Series: A Linguistic Approach to TV Dialogue. Cambridge University Press, Cambridge.
Bordwell, D., 2006. The Way Hollywood Tells It: Story and style in modern movies. University of California Press, Berkeley and Los Angeles.
Bordwell, D., 2007. Poetics of Cinema. Routledge, London and New York.
Corbetta, M., Shulman, G.L., 2002. Control of goal-directed and stimulus-driven attention in the brain. Nature Rev. Neurosci. 3 (3), 201–215.
Flowerdew, J., Mahlberg, M. (Eds.), 2009. Lexical Cohesion and Corpus Linguistic. John Benjamins, Amsterdam.
Friesen, C.K., Kingstone, A., 1998. The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonomic Bull. Rev. 5 (3), 490–495.
Hafri, A., Trueswella, J.C., Strickland, B., 2018. Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. Cognition 175, 36–52.
Halliday, M.A.K., Hasan, R., 1976. Cohesion in English. Longman, London.
Halliday, M.A.K., Matthiessen, C.M.I.M., 2004. An Introduction to Functional Grammar. Edward Arnold, London.
Hasan, R., 1984. Coherence and cohesive harmony. In: Flood, J. (Ed.), Understanding reading comprehension: cognition, language, and the structure of prose. International Reading Association, Newark, Delaware, pp. 181–219.
Hoffmann, C.R., 2012. Cohesive Profiling: Meaning and Interaction in Personal Weblogs. John Benjamins, Amsterdam.
Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Machine Intell. 20 (11), 1254–1259.
Janney, R.W., 2010. Film discourse cohesion, in C.R. Hoffmann, ed., 'Narrative Revisited. Telling a story in the age of new media', number 199 in 'Pragmatics and Beyond', John Benjamins, Amsterdam, pp. 245–266.
Janney, R.W., 2012. Pragmatics and cinematic discourse. Lodz Papers Pragmat. 8 (1), 85–114.
Kluss, T., Bateman, J., Preußer, H.-P., Schill, K., 2016. Exploring the role of narrative contextualization in film interpretation: issues and challenges for eye-tracking methodology. In: Reinhard, C.D., Olson, C.J. (Eds.), Making Sense of Cinema: empirical studies into film spectators and spectatorship. Bloomsbury Academic, New York and London, pp. 257–284.
Kurby, C.A., Zacks, J.M., 2008. Segmentation in the perception and memory of events. Trends Cognitive Sci. 12, 72–79.
Loschky, L.C., Larson, A.M., Magliano, J.P., Smith, T.J., 2015a. What Would Jaws Do? The Tyranny of Film and the Relationship between Gaze and Higher-Level Narrative Film Comprehension. PLoS One 10 (11), e142474.
Loschky, L.C., Larson, A.M., Magliano, J.P., Smith, T.J., 2015b. What would jaws do? the tyranny of film and the relationship between gaze and higher-level narrative film comprehension. PLoS One 10 (11), 1–23.
Maienborn, C., 2019. Event semantics. In Maienborn, C., Heusinger, K., Portner, P. (Eds.), 'Semantics', Walter de Gruyter.
Martin, J.R., 1992. English text: systems and structure. Benjamins, Amsterdam.
Paindaveine, D., 2008. A canonical definition of shape. Statistics Probab. Lett. 78 (14), 2240–2247.
Palmer, R., 1989. Bakhtinian translinguistics and film criticism: the dialogical image? In: Palmer, R. (Ed.), The cinematic text: methods and approaches. AMS Press, New York, pp. 303–341.
Piazza, R., Bednarek, M., Rossi, F. (Eds.), 2011. Telecinematic Discourse: Approaches to the language of films and television series. John Benjamins, Amsterdam.
Radvansky, G.A., Zacks, J.M., 2017. Event boundaries in memory and cognition. Current Opin. Behav. Sci. 17, 133–140.
Royce, T.D., 1998. Synergy on the page: exploring intersemiotic complementarity in page-based multimodal text. Japan Assoc. Systemic Functional Linguistics (JASFL) Occasional Papers 1 (1), 25–49.
Schubert, C., 2017. Discourse and Cohesion. In: Hoffmann, C.R., Bublitz, W. (Eds.), Pragmatics of Social Media', number 11 in 'Handbooks of Pragmatics (HOPS). De Gruyter Mouton, Berlin, pp. 317–344.
Schütt, H.H., Rothkegel, L.O.M., Trukenbrod, H.A., Engbert, R., Wichmann, F.A., 2019. Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. J. Vision 19 (3), 1. https://doi.org/10.1167/19.3.1.
Simons, D., Rensink, R., 2005. Change blindness: Past, present, and future. Trends Cognitive Sci. 9 (1), 16–20.
Smith, T.J., 2012. The attentional theory of cinematic continuity. Projections 6 (1), 1–27.
Stainbrook, E., 2016. A little cohesion between friends; or, we're just exploring our textuality: reconciling cohesion in written language and visual language. In: Cohn, N. (Ed.), Visual Narrative Reader. Bloomsbury, London and New York, pp. 129–157.
Sukthanker, R., Poria, S., Cambria, E., Thirunavukarasu, R., 2020. Anaphora and coreference resolution: A review. Informat. Fusion 59, 139–162.
Tanskanen, S.-K., 2006. Collaborating towards Coherence: Lexical Cohesion in English Discourse. John Benjamins, Amsterdam.
Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M., 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev 113 (4), 766–786.
Tseng, C., 2013. Cohesion in Film: Tracking Film Elements. Palgrave Macmillan, Basingstoke.
Tseng, C., Laubrock, J., Pflaeging, J., 2018. Character developments in comics and graphic novels: A systematic analytical scheme. In: Alexandra Dunst, J.L., Wildfeuer, J., Dunst, A. (Eds.), Empirical Comics Research: Digital, Multimodal, and Cognitive Methods. Routledge, London.

van Leeuwen, T., 1991. Conjunctive structure in documentary film and television. Continuum: J. Media Cultural Stud. 5 (1), 76–114.
van Leeuwen, T., 2005. Introducing social semiotics. Routledge, London.
Wildfeuer, J., 2014. Film Discourse Interpretation. Towards a New Paradigm for Multimodal Film Analysis, Routledge Studies in Multimodality, Routledge, London and New York.
Wolfe, J.M., 1994. Guided search 2.0 a revised model of visual search. Psychon. Bull. Rev. 1 (2), 202–238.
Zacks, J.M., 2010. How we organize our experience into events. Psychol. Sci. Agenda 24.
Zacks, J.M., Magliano, J.P., 2011. Film, Narrative and Cognitive Neuroscience. In: Bacci, F., Melcher, D.P. (Eds.), Art and the Senses. Oxford University Press, Oxford and New York, pp. 435–454.
Zacks, J.M., Speer, N.K., Reynolds, J.R., 2009. Segmentation in Reading and Film Comprehension. J. Exp. Psychol. 138, 307–327.
Zacks, J.M., Speer, N., Swallow, K., Braver, T., Reynolds, J., 2007. Event perception: a mind/brain prespective. Psychol. Bull. 133, 273–293.
Zwaan, R.A., Radvansky, G.A., 1998. Situation Models in Language Comprehension and Memory. Psychol. Bull. 123 (2), 162–185.