Tobias Klauk, Tilmann Köppe, and Thomas Weskott

Empirical Correlates of Narrative Closure

This paper presents an experimental investigation of the narratological concept of *narrative closure*. While narrative closure is a well-studied phenomenon in contemporary narratology, it still lacks a serious empirical foundation. In order to fill that lacuna, we performed a controlled rating experiment aimed at validating some of the properties of narrative closure proposed in the narratological literature. Our results suggest that narrative closure is closely related to two connected properties: to the completeness of the text and to questions left open by the text.

1. Introduction

While narrative closure is typically recognized as an important feature of narrative, narratological literature has reached no consensus as to how to define narrative closure (see e.g. Herrnstein Smith 1968; Kermode 1967 and 1978; Miller 1981; Brooks 1984, esp. 19-22; Brewer 1985, esp. 186; Holland 2009, 164-170; Torgovnick 1981; Branigan 1992, 20; Krings 2004; Abbott 2008, 56-66). One can, however, find a core which seems to be common to the rather diverse remarks and accounts: Narrative closure obtains if informed readers have the impression that the plot of a narrative has ended. This impression might or might not coincide with the actual ending of a text. Narrative closure is, therefore, a reception phenomenon. However, while readers' impressions should in principle be a rewarding object for empirical studies, a central problem for the empirical study of narrative closure is that closure cannot be measured directly, i.e., there is no scale of measurement which would lend itself to the representation of (perceived vs. 'real') closure, as is the case e.g. for psychophysical notions like *loudness*. The present study therefore aims to find empirical correlates of properties that have been argued to be constitutive of narrative closure.

1.1 Empirical Accounts

The phenomenon has, to the best of our knowledge, almost never been empirically investigated: Susan Lohafer (2003) asked 180 readers (including herself) to mark those sentences of stories which in their opinion could end the story. For each reader the five choices nearest to the actual end of the story were recorded. Sentences with high agreement she called "preclosure" points. She then took these findings in order to look for signals which might have prompted the closure impression.

William F. Brewer (1996) presented two stories to readers, each of which was given in three versions, one with a good ending, one with a bad ending, and one with a so-called didactic ending, i.e. an additional sentence concerning the weather or the like which readers were supposed to be able to understand as hints to the outcome of the story, a phenomenon first noted by Victor Shklovsky under the name "false endings" (Shklovsky 1929, 56; Brewer, referring to a different translation, speaks of "illusory endings", Brewer 1996, 265). Readers were then asked to rate the stories concerning overall liking, 'storyness' (i.e. the extent to which a text is taken to constitute a story), outcome liking, completeness, and arrangement. Brewer, without discussing narrative closure explicitly, reports that readers rated stories with bad endings as less complete than stories with good endings, while didactic endings came out somewhere in between. While Lohafer's study lacks empirical rigidity, Brewer touches our subject only in passing. He does not explicitly consider stories without closure, but asks for four criteria (completeness, overall liking, outcome liking, story rating) which have been named in connection with closure, and considers "illusory [i.e. false] endings". Since Brewer only used two texts, this leaves open the possibility that the effects are due to peculiarities of those texts. However, we can learn from Brewer's study that (a) one should consider more than one criterion for closure, (b) one should consider a statistically interesting number of texts to warrant a more robust generalization beyond the sample of texts investigated, (c) one has to be on the lookout for accidental "illusory endings" in items which produce closure where there is supposed to be none, and (d) recipients react to the difference between good and bad story endings. We come back to these points in the description of our experimental design.

1.2 Narratological Accounts

In the narratological literature, a number of quite different aspects are taken to be at the core of the closure phenomenon. In order to find empirical correlates for narrative closure, one cannot therefore rely on narratological consensus; moreover, even the distinction between describing the phenomenon and giving criteria for its occurrence is often blurred in the narratological literature.

Before giving our summary of what we take to be the main accounts of closure, a brief precautionary remark is in order. Some accounts see emotional reader reactions at the heart of the closure phenomenon (see Habermas / Berger 2011, 208; Velleman 2003, with reference to Kermode 1967; see also Miall / Kuiken 2002, 228). Without wanting to take a stance on how central emotional reactions are for narrative closure, we concentrate here on cognitive aspects only. One reason for this restriction is practical. Emotive accounts of closure are typically even vaguer than cognitive accounts. Applying some specifications, however, can help to turn them into empirically more tractable cognitive accounts. We mention but one example. According to Anz,

> literary texts produce suspense by stimulating in the reader […] the experience of
> shortage, which in turn generates the desire to eliminate the shortage. […] The
> energy of desire dissolves, when it reaches its goal or ultimately fails in reaching
> its goal. (Anz 2002, 168, our translation)

Anz talks here about the desires of the reader, but it is difficult to understand what exactly has to happen in order for a text to have narrative closure. We can, however, read the quote in a more cognitive way: texts can generate certain expectations in readers. Texts with narrative closure fulfill these expectations, or else make clear that they do not get fulfilled, while texts without narrative closure do not fulfill the expectations but leave the expectations intact. We come back to such a cognitive account below.

Cognitive accounts of narrative closure can roughly be divided into seven categories, depending on which aspect they consider to be central for narrative closure. Many accounts fall into more than one category.

The *first category* contains accounts which link narrative closure to the *completeness* of a text. Noël Carroll's account is a clear example: "Closure yields a feeling of completeness." (Carroll 2007, 2) Gerald Prince expresses the same thought in slightly different terms, when he speaks of "[t]he feeling of wholeness which a narrative conveys […]" (Prince 1982, 154). Edward Branigan puts the point like this:

> A narrative ends when its cause and effect chains are judged to be totally delineated.
> There is a reversibility in that the ending situation can be traced back to the begin-
> ning; or, to state it another way, the ending is seemingly entailed by the beginning.
> This is the feature of narrative often referred to as *closure*. (Branigan 1992, 20)

Similar thoughts can be found in Eldrige (2007, 68), Richter (1975, 5f.) and Herrnstein Smith: "[Closure] reinforces the feeling of finality, completion, and composure which we value in all works of art." (Herrnstein Smith 1968, 36) Barbara Herrnstein Smith is concerned with poetry. Her book is a classic also of the narratological debate, but one could suspect that closure in poetry might work differently from closure in prose texts. At this point, however, we are merely collecting candidates for prompts, so there is no reason to sort out Herrnstein Smith's ideas just because they were introduced concerning poetry.

A *second category* contains accounts which connect closure to the *fulfillment of reader expectations*. Frank Kermode notices that "[…] certain expectations have been created in him [the reader], and ought to be satisfied." (Kermode 1978, 145) In the same vein, D.A. Miller writes that

> [n]arrative proceeds toward, or regresses from, what it seeks or seems most to
> prize, but it is never identical to it. To designate the presence of what is sought or
> prized is to signal the termination of narrative – or at least, the displacement of
> narrative onto other concerns. (Miller 1981, 272)

In the quote, narrative closure is described without explicitly bringing in the reader. Nonetheless Miller's account is best understood as being concerned with a reception phenomenon: the narrative does not seek or prize, but readers do. The same thought is expressed in slightly different terms by Norman Friedman:

> Thus it is an important formal feature of any good plot to move us to anticipate
> certain things; to mislead us into expecting the wrong things; but to induce us to
> believe, upon looking back, that the way things actually turned out, however sur-
> prising, was nevertheless adequately prepared for and is the only appropriate out-
> come. (Friedman 1975, 69)

Finally, fulfillment of reader expectations has also been named *completion*: "For me to have a satisfying literary experience, I have to be able to bring the plot to closure, and that means completing my expectations." (Holland 2009, 165)

A *third category* of accounts predicts that texts with narrative closure are preferred by readers – they are *liked better*. As Kermode puts it, "[…] we like fictions to end decisively" (Kermode 1978, 154). In the quote given above, Herrnstein Smith also hints at this preference when she remarks that "[closure] reinforces the feeling of finality, completion, and composure *which we value in all works of art*" (Herrnstein Smith 1968, 36, our emphasis). These remarks leave open whether Kermode and Herrnstein Smith think that readers will like whole texts with narrative closure better, or just their endings. The third category therefore has to be taken with caution. This is further revealed by remarks made by David H. Richter linking narrative closure to "a satisfactory ending" on the one hand but also mentioning "balanced irresolutions" in the very same passage which he takes to be explicitly "open-ended" – they don't have narrative closure (Richter 1975, 5). (Richter actually takes closure to depend on the completeness of the text.)

A *fourth category* has it that narrative closure obtains when readers have *no more relevant questions*. Carroll claims that "[c]losure then transpires when all of the questions that have been saliently posed by the narrative get answered" (Carroll 2007, 4). In fact, Carroll explicitly links up the demand for completeness to the answering of questions: "The impression of completeness that makes for closure derives from our estimation, albeit usually tacit, that all our pressing questions regarding the storyworld have been answered." (Ibid., 5) Other accounts understand the fulfillment of expectations in terms of answered questions. When Seymour Chatman writes that "[t]he last event of a narrative may answer all our questions […] which we can often anticipate" (Chatman 1993, 21), one can read the condition of anticipation as expressing what other authors have called reader expectations. Notice, however, that questions and expectations have also been taken to constitute two different kinds of closure. Thus Porter H. Abbott speaks of "closure at the level of expectations" (Abbott 2008, 58) and "closure at the level of questions" (ibid., 60).

Most accounts fall into one of these four categories. The field of narratological claims concerning narrative closure is richer, though. For the next three categories we found only one proponent each. We include them since the ideas in them are clearly expressed and allow for a clear prediction.

Category five is, so to speak, the counterpart to no further questions. Prince claims that at the end of texts without narrative closure, readers *want to know how the text continues*. Having read a text with narrative closure, "[w]e feel that matters are perfectly rounded and that no event preceding or following the sequence of events recounted can be narratively important" (Prince 1982, 153). We take this to mean that, when confronted with a text with narrative closure, readers lack the urge to hear about further events which could shed more light on the events depicted in the text.

Category six sees *soundness of the text* at the heart of narrative closure. Herrnstein Smith, whom we already quoted twice, also expresses this idea, when she writes that

> [closure] gives ultimately unity and coherence to the reader's experience of the poem by providing a point from which all the preceding elements may be viewed comprehensively and their relations be grasped as part of a significant design. (Herrnstein Smith 1968, 36)

Category seven represents the idea that a narrative exhibits closure if and only if it is a *story*. The idea goes back to Brewer who maintains that "[…] narratives […] without a critical event or with no resolution will not be called stories", and "[…] narratives with a significant event and resolution (curiosity discourse structure) will be called stories, whereas narratives without one or the other will not" (Brewer 1985, 172).

In sum, the narratological literature on narrative closure can be taken to supply us with (at least) seven theoretical conceptions of narrative closure which differ in the aspects they put at the center of narrative closure: completeness, fulfillment of reader expectations, the answering of relevant questions, the wish to know what happens after the end of the text, soundness of the text, 'storyness' of the text, and the liking of text or ending. These categories are the starting points for our experiment.

2. Experimental Evidence

In order to put the seven theoretical conceptions of narrative closure to an empirical test under controlled experimental conditions and to quantify their empirical coverage, we designed a rating study in which we systematically varied the narrative closure properties of experimental texts and asked a group of expert participants to judge these manipulated texts numerically on a number of different judgment scales, among them control scales as well as those scales we considered critical to narrative closure. Our goal was to get a first start towards a notion of narrative closure that is empirically valid and to explore what the defining components of narrative closure are. To that aim, we performed an experiment in which participants rated experimental texts along 12 scales under four different conditions of the experimental factor CLOSURE: the experimental texts either had a positive ending, or a negative ending, or an ending that delays the actual ending, or they lacked a proper ending. We predicted that, if one of the 12 scales on which participants had to rate the text is related to narrative closure (i.e., the scale *indexes* a narrative closure-related property), then the mean ratings on this scale should not differ between the positive and the negative ending, since both constitute cases of narrative closure in the most general sense. The delayed and the open ending, however, should show mean ratings that differ from those for the positive and the negative condition, since both constitute cases of violations of the general expectation that a story is 'closed'. Thus, our criterion for a narrative closure-related scale was that the mean ratings

on a given scale do not differ between the positive and the negative ending in terms of inferential statistics, i.e. there should be no significant difference between those conditions, while, at the same time, the scale under consideration should show a significant difference between these two conditions and the delayed one, and the open one. In what follows, we will describe the method employed in this experiment in more detail.

2.1 Method

Participants: The participants in our experiment were 24 undergraduate and graduate students of literary science at the Department of German Studies at the University of Göttingen, 21 females, 2 males and 1 unspecified, age ranging from 23 to 34, mean=26.3, StdDev=2.15. All of them were native speakers of German. They can, given their amount of exposure to literary texts during their studies, surely be called experts on texts and their properties.

    Materials: We constructed 24 experimental texts consisting of three sentences each. These texts had the same overall discourse structure: after a short header introducing an event to which readers attribute a certain prototypical structure (a wedding, a visit to a concert, etc.), the first sentence mentioned a prototypical subevent initiating the event sequence. The second sentence contained information that might lead readers to expect a deviation from the prototypical course of (sub)events as, for example, rain on the day of the wedding which was to be celebrated in the garden. The third sentence contained the experimental manipulation pertaining to narrative closure. It exhibited either a positive (1.3.positive), a negative (1.3.negative), a delayed (1.3.delayed), or an open ending (1.3.open). To give one example of an experimental text, in a rough English translation of the German original:[1]

(1.0)	*The Wedding Day*
(1.1)	The whole family had been planning the wedding party in the garden for quite some time.
(1.2)	On the day of the wedding, it started to rain heavily.
(1.3.positive)	Thanks to the father of the bride, an alternative venue could be found.
(1.3.negative)	The wedding had to be cancelled.
(1.3.delayed)	The father of the bride promised to look for an alternative venue.
(1.3.open)	The garden was in full bloom.

As is obvious from the example, the first two sentences were always constant across the four CLOSURE conditions. Thus, any differences in the judgments

between these conditions have to be attributed to the difference in closure properties exhibited by the last sentence.

We chose this specific example from our list of items here to showcase our awareness for a potential difficulty already mentioned in connection with Brewer (1996). Expert readers are trained to spot 'false endings', i.e. to treat seemingly unconnected events as somehow contributing to the story. (1.3.open) looks like it could be such a case, but in fact is not: readers reacted to the difference between (1.3.positive) and (1.3.open).

We took care that all 24 texts had approximately the same length; the length of the sentences constituting the texts varied between four and 19 words, the latter being an outlier – the mean length was 10 words (*SD*=3.36*)*. The overall length of the texts ranged from 23 to 51 words (*x*=33.4, *SD*=6.34). In contrast to Brewer (1996) the passages were held short in order to be able to control differences in style, wording, anaphorical references, etc. There were no filler texts.

We assessed the overall coherence of 16 texts in the four CLOSURE conditions in a pre-test in which naive participants had to judge the coherence of the last sentence of the text, given the first two, on a 7-point scale ranging from 1 ("completely incoherent") to 7 ("completely coherent"). Participants in this pre-test (*n*=48) judged the texts in the CLOSURE conditions (positive), (negative), and (delayed) to be of approximately the same coherence (6.15 vs. 6.22 vs. 5.96), while only the (open) condition showed a marked decrease in coherence (3.22). We ran the pretest on a total of 24 items of which, for our main experiment, eight were replaced for reasons like stylistic similarity to other items, or the possibility to read sentence three as a 'false ending'. Also, with a further experiment in mind, we wanted to allow for the possibility to understand the text if the sentences were presented in a different order. Thus none of the reasons for replacing an item rested on the criterion of coherence.

Each text was followed by 12 sentences used as judgment prompts and describing properties of the texts (see Kotovych et al. 2011 for a similar design and procedure). The task of the participants was to read each text carefully, and then give a judgment on a 7-point scale whether they thought the description each of the 12 prompts gives is appropriate, given the text, with the scale ranging from 1 ("description is completely inappropriate") to 7 ("description is completely appropriate"). The 12 prompts were the same for all 24 items and pertained to a range of features of text comprehension, (hypothetical) correlates narrative closure being among them (s. below); the words printed in **bold type** will be used as labels for the 12 different judgment prompts:

(P01) **comprehensibility**: The text was comprehensible.
(P02) **completeness**:   The text is complete.
(P03) **chance**:   There was chance involved in the text (in what the text describes).
(P04) **prudence**:   The protagonists in the text acted prudent.
(P05) **textliking**:   I liked the text.

(P06) **expectation**:	The end of the text was according to my expectations.
(P07) **anticipation**:	The protagonists in the text acted foresightfully.
(P08) **endliking**:	I liked the end of the text.
(P09) **soundness**:	The sequence of events described in the text is sound.
(P10) **continuation**:	I would like to know how the text continues.
(P11) **storyness**:	The text tells a story.
(P12) **no question open**:	I don't have any further question with respect to the text.

To control for possible order effects of the 12 judgment prompts, half of the participants received them in the order (S1-S12), and the other half in the reverse order, thereby yielding a further factor P-ORDER (for "prompt order"). The judgment scales were given by a numerical scale, with the end points labeled as "completely appropriate" (7) and "completely inappropriate" (1), respectively.

According to the reasoning laid out in section 1 above, there are certain judgment prompts that should prove to be sensitive to the manipulation of the factor CLOSURE, i.e. their mean judgments should show differences depending on the CLOSURE condition. These are the prompts extracted from the seven categories of narratological closure accounts. For example, the prompt *completeness* (P02) should obviously produce different means for the CLOSURE conditions (positive) and (negative) which are instantiated by texts that are complete in the sense that they "have an ending", and for the CLOSURE conditions (delayed) and (open) which do not instantiate that property. Further prompts that should show a reaction to the CLOSURE manipulation were, by hypothesis, *expectation* (P06), *soundness* (P09), *continuation* (P10), *storyness* (P11), and *no further questions* (P12). *Textliking* (P05) and *endliking* (P08) both go back to category three. Since the narratological literature gives us no hint whether to expect a better liking of the whole text or just of its end, we included both.

Given that we were interested in the question which of these eight properties should be taken as the most reliable empirical correlate of the narrative closure property, we formulated a criterion which a reliable correlate of narrative closure should satisfy:

Criterion (C): A property of narrative closure indexed by the judgment means for prompt *P* is a reliable correlate of narrative closure if and only if the mean ratings for *P* do not differ significantly between CLOSURE conditions (positive) and (negative), and the mean ratings between (positive) and (delayed) do differ significantly: (delayed) should be rated lower than (positive).

The reasoning behind this criterion is simply that any (hypothesized) correlate of narrative closure that does not reliably discriminate between (positive) ending and (delayed) ending is empirically inadequate; and, moreover, that a correlate that does discriminate between (positive) and (negative) ending is equally inadequate, since both conditions are, intuitively, instances of narrative closure. Note

that, since condition (open) is, by hypothesis, 'less closed' than (delayed), it follows by transitivity that in order to satisfy (C), a prompt also has to discriminate between (positive) and (open). We also feared that the strangeness of (open) might affect readers' judgments. Had we relied on the contrast between (positive) and (open) alone, we could not have been sure if effects were due to the strangeness of the (open)-endings, or due to missing closure.

The remaining prompts were used as controls for a number of different properties of the text in the four conditions: *comprehensibility* (P01) checked for potential differences in comprehensibility of the items in the four conditions; *chance* (P03) was a benchmarking prompt to check for inattentive or confused participants: it should show no effect of CLOSURE whatsoever, since its content is completely orthogonal to our experimental manipulation. In order to distract participants somewhat from the objective of our study, we had two prompts that asked to judge properties of the protagonists of the story, *prudence* (P04) and *anticipation* (P07).

We created four experimental lists according to a latin-square design, i.e. from the four experimental CLOSURE versions of each text such that each list contained each item in one CLOSURE condition only, and that each list contained an equal number of items per CLOSURE condition. The order of items was randomized separately for each list. We created four further lists by inverting the order of the four original lists, thereby creating a further between-subject pseudo-factor I-ORDER (for "item order") to check for possible ordering effects. Participants were assigned randomly to one of the eight resulting lists.

Procedure: The experimental task was administered as a paper-and-pencil judgment test. Each participant was handed a small booklet containing inquiries into a few sociographic variables (age, gender, subject of study, handedness), an instruction, and the experimental items plus the prompts. The instruction asked participants to judge the prompt sentences on the 7-point scale and gave an example for two kinds of judgments given a text which should illustrate the nature of the task without prescribing any kind of response behavior. Moreover, participants were asked to use the full range of the scale, not to omit any texts of prompts, and not to backtrack or revise their judgments. Each text was followed by the 12 prompts, one line each; each prompt was followed by its respective numeric scale, with labeled end points. Participants were tested as a group, but worked through the booklets independently; in addition, care was taken that no two neighboring participants had the same list version of the questionnaire.

Design and Predictions: The experimental design was one-factorial, with the four-level factor CLOSURE which was realized within-participants and within-items. In addition, we added a two-level between-subject and within-items control factor P-ORDER pertaining to the order of prompt sentences after the experimental texts, and a further two-level between-subject and within-items control factor I-ORDER checking for possible effects of order of the experimental texts within the questionnaire.

Given the exploratory nature of the experiment, there was no clear-cut prediction as to the different empirical correlates. However, we applied Criterion

(*C*) to identify properties of texts (as indexed by the prompts) which should correlate with narrative closure. Given the theoretical conceptions of narrative closure described in section 1, we surmised that the prompts *completeness* (P02), *expectation* (P06), *soundness* (P09), *continuation* (P10), *storyness* (P11), and *no further questions* (P12) should be among the candidates for satisfying (*C*). For *textliking* (P05) and *endliking* (P06), there were no predictions, since we did not know which of the two were supported by the narratological accounts. Also, while the narratological literature made us expect that one of the two would correlate with closure, the study of Brewer (1996) predicted a decrease between the CLOSURE conditions from (positive) to (negative) for *endliking*. We did not test these predictions in the inferential statistics, but rather only checked the descriptive values for possible quirks in the materials, or any kind of unexpected behavior on the part of our participants.

2.2 Results

Analysis software: For the inferential statistic analysis, we used the *lmer* program of the *lme4* package (Bates / Maechler / Bolker / Walker 2015, version 1.1-9) for the *R* software for statistical computing (version 3.1.1, R Core Team, 2014).

Descriptive Statistics: The descriptive values (means and standard deviations of the judgments) for each of the prompts are given in Table 1.

	CLOSURE condition			
Prompt Type	(positive)	(negative)	(delayed)	(open)
P01:comprehensibility	6.26 (1.29)	6.35 (1.09)	6.26 (1.18)	4.39 (1.96)
P02:completeness	5.76 (1.74)	5.79 (1.54)	4.99 (1.83)	2.81 (1.65)
P03:chance	4.08 (2.03)	4.44 (1.89)	4.36 (2.07)	4.40 (2.06)
P04:prudence	4.65 (1.70)	3.30 (1.45)	4.28 (1.57)	3.45 (1.36)
P05:textliking	4.10 (1.63)	3.94 (1.59)	3.92 (1.65)	3.24 (1.62)
P06:expectation	4.27 (1.85)	4.10 (1.93)	4.10 (1.86)	1.67 (1.09)
P07:anticipation	4.40 (1.82)	3.45 (1.69)	4.06 (1.71)	3.10 (1.55)
P08:endliking	4.69 (1.64)	3.55 (1.86)	3.76 (1.70)	2.69 (1.79)
P09:soundness	6.01 (1.55)	6.01 (1.40)	5.92 (1.28)	2.77 (1.80)
P10:continuation	2.52 (1.54)	3.10 (1.91)	4.07 (2.19)	3.39 (2.15)
P11:storyness	5.44 (1.53)	5.53 (1.43)	5.34 (1.47)	3.85 (1.82)
P12:no further questions	5.53 (1.81)	5.23 (1.82)	4.14 (2.07)	2.67 (1.88)

Table 1: Mean judgments for the 12 prompts dependent on CLOSURE condition (standard deviations in brackets).

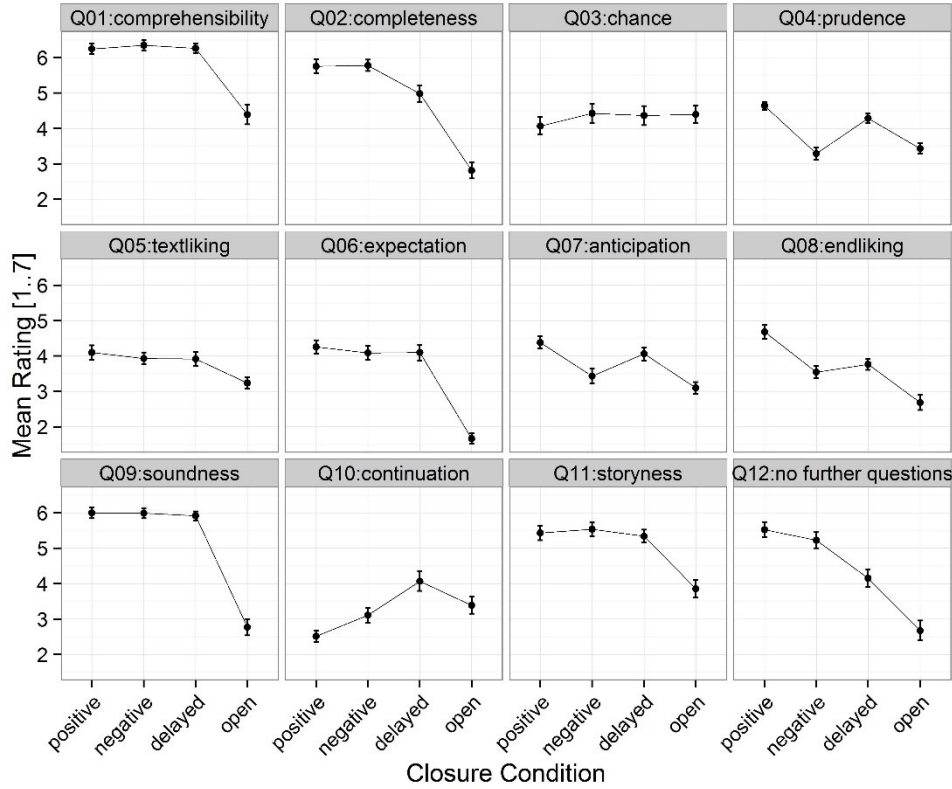These descriptive data are further illustrated by the graph in Figure 1:

Figure 1: Mean judgments for the 12 prompts dependent on CLOSURE condition; error bars depict 1 standard error of the mean.

Before turning to the inferential statistics, we want to point out that our benchmarking prompt, *P03:chance*, which checked for our participant's potential inattentiveness or problems with the task did not show any effect of the CLO-SURE manipulation, which is the outcome to be expected if the participants took the experimental task seriously. Moreover, it shows that participants were aware of and reacted to properties of the experimental text that have nothing to do with narrative closure, since the property this prompt asked them to rate (the contingency of the events described in the text), is completely orthogonal to the CLOSURE manipulation. This result thus makes us confident that the participants of the experiment were unaware of the intent of our study; still, as the mean ratings for the prompts *P02:completeness*, *P10:continuation*, and *P12:no further questions* show, participants were sensitive to the narrative closure properties of the texts, if asked to rate a narrative closure-related prompt statement.

Inferential statistics: In order to assess which of the prompts passed Criterion (*C*) (s. above), we submitted the rating scores for each prompt to a linear mixed model with participants and texts as random factors, and P-ORDER, I-OR-DER, and CLOSURE as fixed factors. Following the suggestions of Barr, Levy, Scheepers, and Tilly (2013), we included the slopes for both random factors. In what follows, we will report *t* values for the effects in question, assuming that effects with $|t| > 2$ are statistically reliable (see Baayen / Davidson / Bates, 2008).

Since neither P-ORDER nor I-ORDER showed any significant main effect, nor contributed to any interactions (all $|t|$s < 1), we will concentrate on the factor CLOSURE. This factor was coded as a treatment contrast in order to compute the planned comparisons for the two pairs of conditions which Criterion (*C*) is about: the comparison of (positive) vs. (negative), for which means should not differ significantly, and (positive) vs. (delayed), which should differ. (The third contrast, i.e. positive vs. open, was of no interest here). The model equation for testing each of the 12 prompt types for Criterion (*C*) thus looked as follows:

lmer(dv ~ closure + (1+closure | participant) + (1+closure | item), data=subset.prompt_type)

This means that we looked at how the dependent variable (dv, here: the rating for a given prompt type) was affected by our experimental manipulation (closure), and also at the interaction of the random error generated by the participant sample and the item sample, as well as their interaction with the experimental manipulation; that is, we included both the random intercepts and the random slopes, as recommended by Barr et al. 2013). Very roughly and glossing over a lot of statistical detail, this means that the experimental effect of our CLOSURE factor is evaluated against the noise coming both from the participant sample, as well as the sample of experimental stories employed.

Let us first go through the results for the control prompts, to which Criterion (*C*) should not apply. Importantly, our control prompt *P03:chance*, as predicted failed to satisfy (*C*), since there was no significant difference between (positive) and (negative) ($|t|$ = 1.01) and also none between (positive) and (delayed) ($|t|$ < 1). The mixed model yielded a similar result for a second control prompt. *P01:comprehensibility* showed no significant effect for the (positive) vs. (negative) contrast, neither did it for (positive) vs. (delayed) (all $|t|$ <1).

The two control prompts that were more related to the content of the narratives, i.e. *P04:prudence* and *P07:anticipation*, showed a similar overall pattern (a 'zigzag') of the mean ratings. For both, there was a significant difference between the (positive) and (negative) condition, thereby violating the first conjunct of *C* (P04: $|t_{a/b}|$ = 6.22; P07: $|t_{a/b}|$ = 3.57).

Neither *P05:textliking* nor *P08:endliking* fulfilled (*C*). *P05:textliking* showed a pattern similar to chance. Here, both *t* values were < 1 for the two tests of (*C*). *P08:endliking* showed a 'zig-zag'-pattern like *P04:prudence* and *P07:anticipation*. As with these prompts, there was a significant difference between the (positive) and (negative) condition, again violating the first conjunct of (*C*) (P08: $|t_{a/b}|$ = 4.43). For P08, the random slopes for the participants had to be removed in order for the model to converge.

Finally, let us have a look at how the remaining prompt types fared with respect to *C*: *P02:completeness*, *P06:expectation*, *P09:soundness*, *P10:continuation*, *P11:storyness* and *P12:no further questions*. All six were hypothesized to be sensitive to the CLOSURE manipulation. Following the suggestion of a reviewer, we included a further criterion (C3) testing – in terms of a planned comparison in the mixed model – for the difference between the positive and the negative condition on the one hand, and the delayed one on the other; this contrast was coded as (1,1-

2,0) across our four conditions, accordingly. Table 2 gives the result of the linear mixed effects model for the two conjuncts of Criterion (*C*) for the four probe types, and, in the last column, the R^2-value for C3, which gives the proportion of variance explained by the C3 contrast:

Probe type	C1 ($|t_{a/b}|$)	C1 ($|t_{a/c}|$)	C (C1 & C2)	marginal R^2 (C3)
P02:completeness	✔ (< 1)	✔ (4.40)	✔	0.3370337
P06:expectation	✔ (< 1)	✘ (< 1)	✘	0.1794534
P09:soundness	✔ (< 1)	✘ (< 1)	✘	0.03716591
P10:continuation	✘ (3.35)	✔ (6.32)	✘	0.07544229
P11:storyness	✔ (< 1)	✘ (< 1)	✘	0.08543131
P12:no further questions	✔ (1.66)	✔ (5.21)	✔	0.2568557

Table 2: satisfaction of the two conjuncts of Criterion (*C*) (subcriteria *C1* and *C2*) for the narrative closure related prompt types, indicated by *t* values.

As the next-to-last column of Table 2 reveals, only the prompt types *P02:completeness* and *P12:no further questions* complied to both subcriteria of (*C*). Furthermore, these prompt types showed the largest amount of variance-explained as indicated by the highest marginal R^2 values for the criterion C3, and thus are the only ones to show the predicted sensitivity to the CLOSURE manipulation employed in our experiment.

3. Discussion

Our experiment reveals two clear empirical correlates of narrative closure. If a text features narrative closure, then readers have the impression that the story is complete, and they have no further questions concerning the text.

3.1 No Questions and Completeness vs. Other Prompts

The pattern of results obtained raises the question whether the two empirical correlates identified have anything in common. Going back to the narratological literature, we can see that at least one way of spelling out the notion of *completeness* is to account for it by the notion of *questions answered*. Recall, for example, Carroll who conceives of completeness in terms of answered questions: "The impression of completeness that makes for closure derives from our estimation, albeit usually tacit, that all our pressing questions have been answered." (Carroll 2007, 5) This would explain why *P02:completeness* and *P12:no further questions* go hand in hand in our experiment. If Carroll is correct, this would also put our findings in perspective in regard to the importance of the prompts. *P02:completeness* is derivative. The impression of completeness is generated by the fact that all relevant questions are answered. While *P02:completeness* is a reliable indicator of closure,

the deeper reason for this is to be found in the reliability of *P12:no further questions*. Notice that *P02:completeness* (like *P12:no further questions*) did not react to the difference between (positive) and (negative), thereby fulfilling the first part of criterion (*C*). This stands in direct contradiction to Brewer's (1996) findings concerning good and bad story endings. Since Brewer used only two texts, we are confident that one should rather generalize from our findings.

It is remarkable that *P10:continuation* has not turned out to be closely connected to *P12:no further questions*. When scanning the narratological literature, we were in fact unsure whether we need to distinguish the wish for a continuation (wanting to know how the text continues) from completeness of answers (still having unanswered questions). The reason for the difference between *P10:continuation* and *P12:no further questions* can be explained if we assume that *P10:continuation* implies *P12:no further questions*, but not the other way around. If a reader wants to know what happens next, she still has unanswered questions. But she might have unanswered questions without actually considering them interesting or important enough to be in need of answering. This assumption can also explain a further detail of the data: Given that *P12:no further questions* scored lowest for (open), i.e. the test subjects had still questions concerning the text, one should expect *P10:continuation* to score highest for condition (open), i.e. test subjects should want to know how the text continues. In fact, ratings for *P10:continuation* are overall relatively low and drop significantly for (open). We assume that test subjects penalized the felt quality of our texts, which were just not interesting enough to raise an interest in continuation. It would be interesting to see whether for more ambitious texts *P10:continuation* might not turn out to be a reliable criterion for narrative closure after all.

Concerning *P06:expectation*, *P09:soundness*, and *P11:storyness*, our findings suggest that fulfillment of expectations, soundness of the text and 'storyness' are not centrally connected to closure. Notice, however, that all three react to the (positive)-(open) difference. Recall that the reason for introducing condition (delayed) was our fear that other effects like the strangeness of (open) might corrupt our findings. *P06:expectation*, *P09:soundness*, and *P11:storyness* seem to justify this fear. We assume that the effect (positive)-(open) for all three prompts is not due to the closure manipulation, but is a product of the strangeness of the open ending. At the same time, we can thereby explain why fulfillment of expectations, soundness of the text and 'storyness' are named in the narratological literature as correlates of narrative closure: violation of reader expectations, and the feeling that the text is not sound and not a story will occur with strange endings. Some of these are open endings without narrative closure, which is why violations of reader expectations, unsoundness and non-'storyness' will co-occur with open endings, although they also occur with other strange endings.

We had no predictions for *P05:textliking* and *P08:endliking*. Although our wording of the prompts was different from Brewer's (1996), we replicated his finding that recipients prefer good endings over bad endings. Also similar to Brewer, the overall liking of the passages was not affected by the difference be-

tween (positive) and (negative). This, however, might again be due to the generally very low score our short texts achieved concerning *P05:textliking*, so that the difference between good and bad endings was probably covered. (Test subjects gave us caring advice how to write better stories.)

For possible follow-up experiments, both *P02:completeness* and *P12:no further questions* have been established to be good candidates for dependent variables indicating closure. Our experimental design however also suggests three caveats: Firstly, we relied on the contrast (positive)-(delayed) because of concerns that for (positive)-(open) other effects like the strangeness of the ending might also play a role. One should, therefore, be cautious when applying *P02:completeness* and *P12:no further questions* to texts with open endings. Secondly, our texts were very short. *P02:completeness* and *P12:no further questions* might be less reliable for longer texts, if these texts are complex enough for readers to diverge in their interests (and that is, in their questions) concerning the text. Thirdly, and closely related to the second point, our design explicitly relies on using schematic knowledge of the test subjects. Remember that the headings already introduced a schematic situation like a wedding, an exam, or a parent-teacher conference. As an anonymous referee for *DIEGESIS* pointed out to us, our results may therefore be specific to stories that have a schema-like structure to them. We feel that almost no text can be understood without using schematic knowledge, which makes us confident that relying on schematic knowledge introduces no important restriction of our findings. Eventual future designs could, however, try to do without reliance on stories with schema-like structures.

3.2 Relation to Narratology

Can our findings be used to say something interesting about narratological theories of narrative closure? Narratologists typically are not interested in just any reader's reaction, but tend to (implicitly) think in terms of informed or even ideal readers. Actual readers (which can be tested in empirical studies) can be warranted in their closure reaction or not, the reaction might be authorized by the text or not. While the conceptual difference between ideal and actual readers forbids any direct transfer of our findings to narratology, we would like to point to two connections:

Firstly, we took care that our test subjects were experts that understood the texts well and, given their competence, did not seem to have any problems with the experimental task. They may therefore be considered close to informed or even ideal readers (see Carroll 2007).

Secondly, while not being directly falsified by our findings, any narratological account which e.g. links closure to 'storyness' must explain why actual readers react differently than such an account would predict.

Therefore, all in all our findings support the general idea of closure accounts like Carroll's (2007) that rely on a question-under-discussion (QUD) approach

to discourse. QUD accounts allow for describing a central property of texts which, as our findings show, is important for narrative closure, namely that texts generate a structured set of questions which they may or may not exhaustively answer, generating narrative closure only in the former case, but not in the latter.[2]

Bibliography

Abbott, Porter H. (2008): *The Cambridge Introduction to Narrative*. Cambridge.

Anz, Thomas (2002): *Literatur und Lust*. München.

Baayen, R. Harald et al. (2008): "Mixed-effects Modeling with Crossed Random Effects for Subjects and Items". In: *Journal of Memory and Language* 59, pp. 390-412.

Barr, Dale J. et al. (2013): "Random Effects Structure for Confirmatory Hypothesis Testing: Keep it Maximal". In: *Journal of Memory and Language* 68, pp. 255–278.

Bates, Douglas et al. (2015): "lme4: Linear Mixed-Effects Models Using 'Eigen' and S4. R package version 1.1-9", URL: http://CRAN.R-project.org/package=lme4 (6.10.2015).

Branigan, Edward (1992): *Narrative Comprehension and Film*. London / New York.

Brewer, William F. (1985): "The Story Schema: Universal and Culture-Specific Properties". In: Olson, David R. et al. (Eds.), *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing*. Cambridge, pp. 167-194.

Brewer, William F. (1996): "Good and Bad Story Endings and Story Completeness". In: Kreuz, Roger J. et al. (Eds.), *Empirical Approaches to Literature and Aesthetics*. Norwood, pp. 261-274.

Brooks, Peter (1984): *Reading for the Plot: Design and Intention in Narrative*. New York.

Carroll, Noël (2007): "Narrative Closure". In: *Philosophical Studies* 135, pp. 1-15.

Chatman, Seymour (1993): *Reading Narrative Fiction*. New York.

Eldridge, Richard (2007): "Das Ende der Erzählung". In: Alex Burri / Wolfgang Huemer (Eds.), *Kunst Denken*. Paderborn, pp. 67-74.

Friedman, Norman (1975): *Form and Meaning in Fiction*. Athens GA.

Habermas, Tilmann / Berger, Nadine (2011): "Retelling Everyday Emotional Events: Condensation, Distancing, and Closure". In: *Cognition and Emotion* 25, pp. 206-219.

Herrnstein Smith, Barbara (1968): *Poetic Closure: A Story of How Poems End*. Chicago / London.

Holland, Norman N. (2009): *Literature and the Brain*. Gainesville.

Kermode, Frank (1967): *The Sense of an Ending: Studies in the Theory of Fiction*. New York.

Kermode, Frank (1978): "Sensing Endings". In: *Nineteenth-Century Fiction* 33, pp. 144-158.

Kotovych, Maria et al. (2011): "Textual Determinants of a Component of Literary Identification". In: *Scientific Study of Literature* 1 (Nr. 2), pp. 260-291.

Krings, Constanze (2004): "Zur Analyse des Erzählanfangs und des Erzählschlusses". In: Wenzel, Peter (Ed.), *Einführung in die Erzähltextanalyse: Kategorien, Modelle, Probleme*. Trier, pp. 163-179.

Lohafer, Susan (2003): *Reading for Storyness. Preclosure Theory, Empirical Poetics, and Culture in the Short Story*. Baltimore / London.

Miall, David S. / Kuiken, Don (2002): "A Feeling for Fiction: Becoming What We Behold". In: *Poetics* 30, pp. 221-241.

Miller, D. A. (1981): *Narrative and its Discontents: Problems of Closure in the Traditional Novel*. Princeton.

Prince, Gerald (1982). *Narratology: The Form and Functioning of Narrative*. Berlin et al.

Richter, David H. (1975): *Fable's End. Completeness and Closure in Rhetorical Fiction*. Chicago.

Shklovsky, Victor (1990): *Theory of Prose*. Champaign / London.

Torgovnik, Marianna (1981): *Closure in the Novel*. Princeton.

Velleman, J. David (2003): "Narrative Explanation". In: *The Philosophical Review* 112, pp. 1-25.

Dr. Tobias Klauk
CRC "Textstrukturen"
University of Göttingen
E-mail: tobias.klauk@zentr.uni-goettingen.de

Prof. Dr. Tilmann Köppe
CRC "Textstrukturen"
University of Göttingen
E-mail: tilmann.koeppe@zentr.uni-goettingen.de

Dr. Thomas Weskott
Seminar für Deutsche Philologie
University of Göttingen
E-mail: thomas.weskott@phil.uni-goettingen.de

---

[1] The original materials can be obtained from the authors upon request.
[2] Work on this paper has been supported by the Courant Research Centre "Textstrukturen" at Göttingen University. We would like to thank Norbert Groeben and the anonymous referees for *DIEGESIS* for various helpful suggestions.