




The (In-)Consistency of Literary Concepts Operationalising, Annotating and Detecting Literary Comment

Anna Mareike Weimer¹ 
Florian Barth² 
Tillmann Dönicke² 
Luisa Gödeke¹ 
Hanna Varachkina³ 
Anke Holler¹ 
Caroline Sporleder² 
Benjamin Gittel¹ 

1. Department of German Philology, University of Göttingen , Göttingen, Germany.
2. Göttingen Centre for Digital Humanities, University of Göttingen , Göttingen, Germany.
3. Göttingen State and University Library, University of Göttingen , Göttingen, Germany.

Citation

Anna Mareike Weimer, Florian Barth, Tillmann Dönicke, Luisa Gödeke, Hanna Varachkina, Anke Holler, Caroline Sporleder, and Benjamin Gittel (2022). "The (In-)Consistency of Literary Concepts. Operationalising, Annotating and Detecting Literary Comment". In: *Journal of Computational Literary Studies* 1 (1). [10.48694/jcls.90](https://doi.org/10.48694/jcls.90)

Date published 2022-12-16

Date accepted 2022-03-31

Date received 2021-12-21

Keywords

literary theory, narratology, commentary, operationalisation, annotation, supervised machine learning

License

CC BY 4.0 

Reviewers

Martin Eve 

Evelyn Gius 

Note

This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 1st Annual Conference of Computational Literary Studies 2022 at the Technische Universität Darmstadt.

Abstract. This paper explores how both annotation procedures and automatic detection (i.e. classifiers) can be used to assess the consistency of textual literary concepts. We developed an annotation tagset for the ‘literary comment’ – a frequently used but rarely defined concept – and its subtypes (interpretative comment, attitude comment and metanarrative/metafictional comment) and trained a multi-output and a binary classifier. The multi-output classifier shows F-scores of 28% for attitude comment, 36% for interpretative comment and 48% for meta comment, whereas the binary classifier achieves F-scores up to 59%. Crucially, both our annotation and the automatic classification struggle with the same subtypes of comment, although annotation and classification follow completely different procedures. Our findings suggest an inconsistency in the overall literary concept ‘comment’ and most prominently the subtypes ‘attitude comment’ and ‘interpretative comment’. As a best-practice-example, our approach illustrates that the contribution of Digital Humanities to Literary Studies may go beyond the automatic recognition of literary phenomena.

1. Introduction

While Computational Literary Studies received much attention in recent years, the potential for collaboration between traditional Literary Studies and the Digital Humanities has not yet been fully explored. Arguments about the benefits of digital methods – often framed as promises for the future of Literary Studies – flourish, including: (1) a systematic application of concepts developed within Literary Studies (in what follows: ‘literary concepts’) in the process of annotation leads to refining their definitions (cf. Gius and Jacke 2015; Gius and Jacke 2017); (2) the use of quantitative methods may lead to “new forms of evidence” for literary phenomena and to a “scientification” of Literary Studies (Jockers 2013, 5–10, here: 8); (3) insofar as automatic recognition of literary phenomena succeeds, a large number of examples can be readily retrieved and

submitted to qualitative analysis (cf. Piper et al. 2021); (4) if an automatic recognition of literary phenomena in representative diachronic corpora is successful, it is possible to model developments in literary history (cf. Underwood 2016; Underwood 2019) and justify claims about generic literary entities like ‘the novel’ (cf. Piper 2018, xi). In addition – so we will argue – Digital Humanities could help to examine which literary concepts are useful for quantitative empirical research, thus potentially reducing the abundance of literary concepts. As has been shown in recent years, however, by no means all attempts to operationalise and automatically detect such literary concepts were successful. Some seem to resist operationalisation and/or automatic detection (cf. Herrmann et al. 2015; Willand et al. 2020) and in some cases the reason for this may be the inconsistency of the literary concept itself.

By the ‘consistency of a concept’, we mean that (a) comparatively homogeneous phenomena fall under it and (b) that the concept is methodologically guiding in the sense that these phenomena are intersubjectively and automatically recognisable. Inconsistent concepts, on the other hand, describe comparatively heterogeneous phenomena, which cause hardly or not at all surmountable difficulties when trying to recognise them intersubjectively and/or automatically.

We will presume that a consistency study is feasible and demonstrate this using a concrete example: the ‘literary comment’, which lends itself very well to such an approach. Comments can be used to clarify a narrator’s / character’s attitude, to steer the reader’s attention, interpret or explain plot elements, reflect about the real world, the narration or the literary work (Gittel 2022), or signal an “overt-narrator” (Chatman 1980). As intuitively easy to understand as ‘comment’ may seem at first glance, it nevertheless turns out to be surprisingly imprecise due to sketchy definitions and competing conceptualisations.

In order to explore the consistency of a literary concept it is not sufficient to operationalise, annotate and detect it automatically and evaluate whether annotation and automatic detection succeeded or failed (by measuring inter-annotator agreement or a classifier’s performance). Rather, we seek to explain patterns of the annotation and automation experiments using qualitative and quantitative evidence such as measures of the relation between available training data and classifier-performance, the features found to be predictive for automatic classification as well as assessment and contextualisation of the relevant conceptualisations based on textual examples. Specifically, a consistency study carries out inferences to the best explanation where the inconsistency of a literary concept is a possible hypothesis among others that may explain certain outcomes. We utilise empirical observations from annotation and automation to gain insights into the consistency of the theoretical concept itself.

In the next section, we will scrutinize narratological research on the literary comment (section 2). We then operationalise the notion of ‘comment’ and follow Chatman (1980) in distinguishing subtypes (section 3). We report the results of a collaborative annotation effort (section 4) and an automatic classification (section 5). Finally, we discuss limitations and challenges of consistency studies of literary concepts in general (subsection 6.1), the results of our consistency study for the three types of literary comment (subsection 6.2), and conjectures on the overarching concept of ‘comment’ (subsection 6.3).

2. Theoretical Background

Although the concept of ‘comment’ is known in Literary Studies, it has not yet been systematically considered through the perspective of a consistency study.¹ Rather, the concept ‘comment’ is often used in its commonplace understanding and would benefit from a more detailed analysis. In narratology, a comment usually is associated with the narrator making remarks on what is narrated, that interrupt the narration (cf. Zeller 1997) or with authorial intrusion (cf. Dawson 2016). Its function goes beyond the description of action. Comments explain the meaning of a narrative element, make value judgments, and/or refer to the real world (cf. Prince 2003). The following thoughts are essentially based on the influential contributions of Bonheim 1975 and Chatman 1980. While Bonheim draws attention to structural features of comment, Chatman is interested in the multiplicity of phenomena subsumed under the concept of ‘comment’.

Bonheim examines modes of narrative in accordance with their function, distinguishing between dynamic and static modes on the basis of their temporal constitution of discourse. The comment is treated as a static mode, along with the mode ‘description’, and is thus contrasted with the dynamic modes ‘speech’ and ‘report’. Basically, the modes may overlap, but according to Bonheim, comment is the most autonomous and thus the purest of the modes and is most often found unblended (cf. Bonheim 1975, p. 332). While Bonheim does not define linguistic indicators for what constitutes a comment, he formulates criteria on the text-structural level: A comment must be embedded in a narrative pause and need not be descriptive (e.g. describing the scenery of the narrative).

In a narrative pause (cf. Lahn and Meister 2013, p. 154 drawing on Genette 1994 [1972]), the narrating time exceeds the narrated time such that readers might get the impression the narrated time stops or slows down extremely, although information is provided. However, the concept of ‘narrative pause’ is not unproblematic, since, for now, there is no objective measure for narrating time (beside the word quantity indicator) and the determination of the narrated time may require complex interpretive decisions in individual cases.

Chatman distinguishes four types of explicit comment not as a mode of narrative, but as a quality of sentences or text passages (cf. Chatman 1980).² In the following, we will take a closer look at the four comment types.

Generalisation Chatman defines generalising comments as general truths that can apply not only to the fictional but also the real world. He takes his cue from Booth 1983, who speaks of generalisation as the reinforcement of norms.

(1) Sie [Otilie] ward den Männern vorgestellt und gleich mit besonderer Achtung

1. Terminologically, both the term ‘comment’ (Bonheim 1975) and ‘commentary’ (Chatman 1980) are applied in literary studies. ‘Commentary’ is a term with multiple meanings, often used colloquially in its narratological sense and with other meanings in historical criticism and journalism. In the following, we use ‘comment’ in the article for the sake of uniformity.

2. Chatman makes an additional distinction between implicit and explicit comment. The former includes statements by unreliable narrators and ironic remarks that must be reconstructed by the reader and interpreted from the context. In the following, we focus on the explicit comments and leave out the implicit communication on account of its complexity.

als Gast behandelt. Schönheit ist überall ein gar willkommener Gast. (Goethe 2012 [1809])³

From a linguistic perspective, ‘generalisation’ is an umbrella term for phenomena like genericity and (overt) quantification. Thus, several linguistic markers might be associated with ‘generalisations’. We will come back to problems resulting from this in [section 3](#).

Interpretation The speaker, mostly the narrator, explains the plot pro- or analeptically and provides additional information to help readers correctly understand what is being told. Example (2), the end of E.T.A. Hoffmann’s *Der Sandmann*, illustrates this usage. We provide context to also clarify the function of the narrative pause.

- (2) Als Nathanael mit zerschmettertem Kopf auf dem Steinpflaster lag, war Coppelius im Gewühl verschwunden. - Nach mehreren Jahren will man in einer entfernten Gegend Clara gesehen haben, wie sie mit einem freundlichen Mann, Hand in Hand vor der Türe eines schönen Landhauses saß und vor ihr zwei muntre Knaben spielten. Es wäre daraus zu schließen, dass Clara das ruhige häusliche Glück noch fand, was ihrem heiteren lebenslustigen Sinn zusagte und das ihr der im Innern zerrissene Nathanael niemals hätte gewähren können. (E. T. A. Hoffmann 2012 [1816/17])

In the first sentence of this example, plot is conveyed. The dash indicates a time jump. In the following narrative pause we are given an insight into what happened to Clara after the end of the narration: The narrator offers a description of Clara’s situation. In a third step (the underlined passage), this description is interpreted by the narrator and therefore a comment.

Judgment Evaluative comments formulate the narrator’s judgment reflecting values, norms and beliefs. They are intended to confront the reader with ethical aspects included in the story. Chatman distinguishes between interpretation and judgment only on the basis of the (moral) evaluation underlying the judgment, while interpretation is “relatively value-free” (Chatman 1980, 237).

- (3) Charlotte benutzte des andern Tags auf einem Spaziergang nach derselben Stelle die Gelegenheit, das Gespräch wieder anzuknüpfen, vielleicht in der Überzeugung, daß man einen Vorsatz nicht sicherer abstumpfen kann, als wenn man ihn öfters durchspricht. (Goethe 2012 [1809])

Here, the character’s decision to wear down the partner by repeatedly talking through the controversy is commented on by justifying it with a conviction – or at least the narrator strongly assumes this motivation with Charlotte, as is shown by *vielleicht* ‘perhaps’ which he uses to reflect his own conviction through this comment. This example also contains another type of Chatman’s types of comment: generalisation. This is because the generalisation of what the narrator is convinced of is presented as a universally valid truth.

3. Translations for all examples are provided in [Appendix A](#).

Comment on the Discourse This type of comment, which we will call ‘meta comment’, expresses reflections on the process of writing and/or the existence of the respective work and its fictionality itself.

- (4) Ich verspräche gerne diesem Buche die Liebe der Deutschen. Aber ich fürchte, die einen werden es lesen, wie ein Kompendium, [...] indes die andern gar zu leicht es nehmen, und beide Teile verstehen es nicht. (Hölderlin 2012 [1797])

Meta comment has been extensively studied in other contexts as a category of its own (see for example Fludernik 2003 or Nünning 2005). This includes metanarrative comment and metafictional comment, the latter discussing truth, fictivity and/or fictionality of the respective work.

Given the heterogeneity of phenomena that have been subsumed under the concept of ‘comment’ in narratological research, the question arises, how we may be able to annotate and automatically detect comments in texts. We address this in the next section.

3. Operationalisation

Concepts in Literary Studies including Narratology are often designed from a theoretical point of view and only selectively consider textual examples. Applying them on a larger scale often reveals incompleteness or discrepancies within the theory. Thus, making a literary phenomenon more tangible through annotation requires an iterative process of refinement of the concept utilising complete texts or longer parts of works instead of hand-picked examples (see Gius and Jacke 2017).

The starting point for our operationalisation of the comment are the findings from the previous section: Even if the category ‘literary comment’ seems intuitively coherent and comprehensible, our examination of its conceptualisation revealed that comments are often defined *ex negativo* (see for example Bonheim 1975 or Prince 2003). Interestingly, instead of defining comment, researchers restrict themselves to create open lists of indicators or partial phenomena of comment. Thus, the state of the art seems to suggest that there is no robust concept of ‘comment’, but rather a number of related phenomena that have been subsumed under the overarching concept.

Combining the approaches of Bonheim and Chatman, we assume comment is present if a narrative pause is identifiable (Bonheim) and characteristic features of one of Chatman’s comment types are present in it:

comment := narrative pause AND (interpretative passage OR attitude passage OR meta passage)

By this procedure we exclude blending of the modes to the extent that we do not include comments if they appear linked to dynamic elements, but can thus achieve a higher comparability of the collected data and lower the amount of interpretation required of the annotators.

Let us first look at our approach of detecting a narrative pause. Since these readings are widely unpredictable (section 2), we decided not to pre-determine sentence structures: Our annotation relies completely on intuitive (i.e. form-independent) recognition of

narrative pauses.⁴ This procedure enables us to maintain the explorative character of our narrative pause detection.⁵

As described above, Chatman's types of comments are 'generalisation', 'interpretation', 'judgment' (attitude) and 'commentary on the discourse' (meta comment). In contrast to his informal usage of the term, we understand 'generalisation' as a linguistic phenomenon triggered by e.g. generic terms and quantificational expressions. These can co-occur with any of the subtypes (see e.g. (1) and (2)), but do not constitute a subtype on their own. Since we examine generalisation as a separate category (cf. Gödeke et al. [to appear](#)), three typological manifestations of comment emerge, (i) the attitude comment, (ii) the interpretive comment, and (iii) the meta comment.

(i) Interpretive Comment The interpretive comment offers an interpretation of events within the diegesis. Sometimes it takes the form of an explanation of events. This type of comment can be recognised by the fact that additional information is provided that re-perspectives, interprets or corrects elements of the plot or events within the diegesis. As shown in (2), Clara's situation at the end of the story is interpreted by the narrator.

(ii) Attitude Comment In the attitude comment, an attitude of the speaker (narrator or character) to the diegesis is expressed. By 'attitude', we mean the way in which a speaker views something or feels about something. This includes all objects of the narrative, such as characters, the plot, fictional objects and the fictional world (order) as well as self-references. In (3), presented above, the speaker's attitude towards Charlotte's talking through the argument topic becomes clear.

Here we have made significant changes to Chatman's broad notion of this subtype of comment, which he calls "judgment" and understands as evaluations being based on norms, values and beliefs of the narrator. He uses this criterion as a demarcation to the comment type 'interpretation' which he takes to be "relatively value-free" (Chatman 1980, 237). Since the vagueness of this criterion led to difficulties during the annotation, we decided to annotate the speaker's attitude, as this is more clearly identifiable and the explicit result of the evaluation process. Therefore, we call the subtype 'attitude comment' to make the difference clear.

(iii) Meta Comment The meta comment combines two aspects: metafictional and meta-narrative comment. It reveals the narrator's attitude toward the narrative, its process of creation (narrating) or its truth-status. Since its identification relies on direct mentions of the context and circumstances, in which the respective work of literature was created, we consider meta comment easier for the annotators to identify.

Based on the presented typology of comment, we created a tagset and annotation guidelines.⁶ Accordingly, the tagset for comment includes three subtags: INTERPRETATION, EINSTELLUNG (attitude), and META that correspond to (i), (ii), and (iii). The annotators are supposed to assign these subtags to passages, where a passage can comprise one or

4. This procedure includes the understanding that a narrative pause can also occur in direct speech, which we understand as a narrative structure in itself. This allows us to include comments made by characters and not only those made by the narrator or so-called "authorial insertions" (Dawson 2016).

5. In doing so, our approach differs from, for example, Vauth et al. 2021, who categorise verbal phrases by their eventness from non-event up to change of state.

6. Our annotation guidelines are available at <https://gitlab.gwdg.de/mona/korpus-public>.

Year	Author: Text	#		κ		γ	
		Pa.	Cl.	M.	B.	M.	B.
<i>Training set</i>							
1616	Andreae: <i>Die chymische Hochzeit</i>	47	66	.26	.29	.23	.25
1645	Zesen: <i>Adriatische Rosemund</i>	45	244	.71	.85	.64	.68
1668	Grimmelshausen: <i>Der abenteuerliche Simplicissimus</i>	53	205	.26	.26	.20	.17
1731	Schnabel: <i>Die Insel Felsenburg</i>	73	203	.74	.91	.69	.76
1747	Gellert: <i>Das Leben der schwedischen Gräfin von G.</i>	34	187	.61	.59	.54	.57
1771	LaRoche: <i>Geschichte des Fräuleins von Sternheim</i>	60	282	.33	.33	.39	.38
1797	Hölderlin: <i>Hyperion oder der Eremit in Griechenland</i>	72	313	.41	.76	.46	.51
1802	Novalis: <i>Die Lehrlinge zu Sais</i>	73	400	.61	.71	.54	.52
1809	Goethe: <i>Die Wahlverwandtschaften</i>	138	619	.34	.34	.41	.41
1810	Kleist: <i>Michael Kohlhaas</i>	36	72	.08	.09	.07	.12
1816	Hoffmann: <i>Der Sandmann</i>	37	103	.46	.46	.28	.28
1876	Dahn: <i>Kampf um Rom</i>	43	157	.28	.27	.13	.08
1893	May: <i>Winnetou II</i>	45	79	.55	.68	.60	.65
1898	Fontane: <i>Der Stechlin</i>	54	219	.31	.31	.22	.24
1924	Mann: <i>Der Zauberberg</i>	45	133	.41	.48	.34	.34
1930	Musil: <i>Der Mann ohne Eigenschaften</i>	47	317	.83	.83	.73	.73
1931	Kafka: <i>Der Bau</i>	55	280	.68	.68	.54	.56
		47	205	.46	.52	.41	.43
<i>Test set</i>							
1766	Wieland: <i>Geschichte des Agathon</i>	60	282	.58	.60	–	–
1942	Seghers: <i>Das siebte Kreuz</i>	48	92	.43	.48	–	–

Table 1: For each text, the number of comment passages (Pa.) in the gold standard and the number of clauses (Cl.) overlapping with them, and multi-label (M.) and binary (B.) agreement values in terms of κ and γ . The last row for the training set shows the median counts and the average agreement values.⁷

several clauses. These clauses usually follow one another, but discontinuous annotations are also possible. As for the narrative pause, we do not pre-select any linguistic properties as unique indicators for comment subtags, i.e. the annotation is solely based on a passage's reading and not its form. Comment is a phenomenon that tends to span rather long parts of text. One passage can be labelled with more than one comment subtags. Passages labelled with different subtags can overlap.

4. Corpus and Annotation

Our corpus consists of 19 texts covering the time period from 1616 to 1942. 17 texts serve as training set for the classifiers described in section 5. All six annotators are students with a background in German Philology. In general, the first approximately 200 sentences of each text were annotated by two annotators with the three subtags. Two texts were annotated by all six annotators in order to have a better insight into the feasibility of our approach. We created gold standards for all texts by having 2–3 experts (authors of this paper) collaboratively adjudicate (i.e. review, accept, correct or delete) the initial annotations. Table 1 shows for each text the annotated comment passages and

⁷ We do not show γ for the test texts since the Python package *Pygamma-agreement* (<https://github.com/bootphon/pygamma-agreement>) used for calculation throws a runtime error for 6 annotators.

Subtag	κ	γ
EINSTELLUNG	.44	.50
INTERPRETATION	.26	.28
META	.71	.66

Table 2: Average agreement for subtags.

the number of annotated clauses.⁸ Overall, we observe a median of 47 passages and 205 clauses for comments per text.⁹

To evaluate the annotation, we calculate inter-annotator agreement on clause-level with Fleiss' Kappa (κ , Fleiss 1971) and Mathet's Gamma (γ , Mathet et al. 2015). While κ calculates agreement based on the differences for each clause, γ respects the individual annotated comment passages as units in a continuum, and also partial overlapping passages are compared as units instead of disjointed clauses.¹⁰ We therefore consider that γ better represents the errors made by annotators for a category with rather long passages, and that it measures agreement more adequately. The multi-label values for both scores are based on the agreement between the subtags; binary agreement treats all subtags as a single class (COMMENT). The average binary agreement for κ and γ is moderate (between 0.43 and 0.52; see agreement levels in Landis and Koch 1977). The multi-label agreement is 0.04 lower on average but can still be regarded as moderate.

As hypothesised in section 3, META is easier to annotate since it directly addresses either the way content is mediated or the creation of the respective work. This can be observed when calculating agreement scores directly on the individual subtags as shown in Table 2. The subtag META achieves substantial agreement (0.71 for κ and 0.66 for γ), in contrast, EINSTELLUNG holds moderate values, and INTERPRETATION only achieves a fair agreement ($0.2 < x \leq 0.4$). As pointed out above, especially the distinction between EINSTELLUNG and INTERPRETATION can be difficult, and a decision for only one of both can cause disagreement between the annotators. This effect can be verified when calculating the binary agreement for EINSTELLUNG+INTERPRETATION, which yields a κ of 0.49 and γ of 0.43.

5. Automatic Classification

To gain insights into the consistency of the category 'comment', we employ diverse linguistically available features that we consider to be potentially relevant based on manual inspection of annotated comment passages. In the following, we describe the feature extraction, the classifiers and their evaluation, and then turn to a comprehensive analysis.

8. The number of comment passages and clauses for Goethe's *Die Wahlverwandtschaften* is higher since this was our first annotation, where we annotated the complete first four chapters.

9. We use the median rather than the average because the former is robust against outliers (i.e. texts with an extremely high or low number of comments), and thus better resembles the typical number of comments in a text.

10. The units for γ consist of items, which can be characters, tokens, or clauses. We use clauses since we calculate agreement on clause-level. When calculating γ based on characters or tokens, the result only changes slightly.

5.1 Feature Extraction

We preprocess texts with spaCy,¹¹ using its default tokeniser, part-of-speech (POS) tagger, lemmatiser and sentenciser for German and adding several custom preprocessing components:

- a dictionary-based normaliser that we trained on the German Text Archive¹² to account for spelling variants in older texts;
- the Universal Dependency parser, morphological analyser, clausiser and tense–mood–voice–modality tagger from Dönicke 2020;
- a direct speech tagger that recognises text between opening and closing quotation marks;
- a component that assigns Levin 1995’s categories to verbs and Hundsnurscher and Splett 1982’s categories to adjectives from GermaNet (cf. Hamp and Feldweg 1997);
- the sentiment tagger¹³ from Remus et al. 2010 as well as our own emotion tagger based on the NRC Word-Emotion Associated Lexicon¹⁴ (Mohammad and Turney 2010, 2013), which assign scores for positive/negative sentiment and Ekman 1992’s basic emotions, respectively, to each token.

Unit	Features
clause	root’s dependency relation, root’s POS, preceding/inner/succeeding punctuation, first clause of a sentence?, directed distance to superordinate clause, direct speech?
NP	head’s dependency relation, head’s POS, adpositional?, case, person, number, gender, sentiment, emotion, article’s POS, article’s lemma, quantifier’s POS, quantifier’s type ¹⁵ , adjective’s POS, adjective’s degree, adjective’s GermaNet category, adjective’s sentiment, adjective’s emotion
(composite) verb	main verb’s dependency relation, main verb’s POS, verb form, tense, aspect, mood, voice, modal verb’s lemma, main verb’s GermaNet category, sentiment, emotion, quantifier’s POS, quantifier’s type ¹⁵
free discourse element	dependency relation, POS, at first/middle/last position?

Table 3: Extracted features for different syntactic units.

Inspired by Dönicke 2021’s grammatical feature extraction for discourse segmentation, we extract features clause-wise from the clause, its noun phrases (NPs), the composite verb and free discourse elements (i.e. conjunctions, complementisers, sentential adverbs). Table 3 shows all features. Grammatical features have been found to work well for the identification of discourse segments – which are also a multi-clause-level phenomenon – in German (cf. Dönicke 2021) and might also include useful features for comment identification. For example, we expect verb categories such as tense or mood to be especially useful since a change in those often marks a narrative pause, as in (2). Here,

11. See <https://spacy.io/> (version 2.3.2).

12. See: <https://www.deutschestextarchiv.de/download>.

13. See <https://github.com/Liebeck/spacy-sentiws>.

14. See <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>.

15. We use word lists to classify overt quantifiers with Dönicke et al. 2021’s tagset; except for numerical quantifiers, which we identify by POS (NUM) and/or dependency relation (nummod).

Setting	Development set	#Clauses
split 1	Grimmelshausen (1668), Schnabel (1731)	408
split 2	Mann (1924), Kafka (1931)	413
split 3	Gellert (1747), Fontane (1898)	405

Table 4: Texts in the development set and number of clauses overlapping with comment passages in each split.

the narrated time is interrupted, and the present tense in the first sentence changes to the past tense in the second one.

Punctuation is also integrated as feature. In (4), punctuation marks the direct speech, in which a comment is contained. We also integrate semantic categories for verbs and adjectives. Main verbs like *lesen* ‘read’ and *verstehen* ‘understand’ belong to Levin’s category of cognition, which we assume to be indicative for comments.

Since comment, especially attitude, can be expressed in an emotional manner, we include emotion and sentiment labels as features. (5) shows an excerpt for EINSTELLUNG that is highly expressive due to the usage of so-called “thick concepts”, such as *offen* ‘expansive’ and *wundersam* ‘miraculous’, which “combine evaluation and non-evaluative description” (Väyrynen 2021).

- (5) Wer also ihr [der Natur] Gemüth recht kennen will, muß sie in der Gesellschaft der Dichter suchen, dort ist sie offen und ergießt ihr wundersames Herz EINSTELLUNG.
(Novalis 2012 [1802])

5.2 Classifiers

Since a comment passage spans an open number of clauses, we define a classification task at the clause level: When vectorising a text $D = (c_1, \dots, c_n)$ with n clauses, we construct feature vectors $\vec{c}_1, \dots, \vec{c}_n$ as described in subsection 5.1, which we then concatenate to context-sensitive vectors $X_D = (\vec{x}_1, \dots, \vec{x}_n)$ using a window of three clauses: $\vec{x}_i := \vec{c}_{i-1} \circ \vec{c}_i \circ \vec{c}_{i+1}$. Given \vec{x}_i , the classifier should predict all tags of passages that contain c_i . In a post-processing step, every maximal sequence of clauses with the same tag is combined into a passage, which is, however, not relevant for the evaluation, see subsection 5.3.

From our training set, we remove two texts as development set. To alleviate the impact of the split, we perform our experiments for three different splits as shown in Table 4. Since the median count of comment clauses per text is 205 (see Table 1), we take two texts with a total number of around 410 comment clauses in each split. Furthermore, split 1 uses two early texts, split 2 uses two late texts, and split 3 uses an earlier and a later text as development set.

In each split, we train (1) a multi-output classifier that consists of three independent binary classifiers (one for every subtag), and (2) a binary classifier that only distinguishes comment (any subtags) from non-comment (no subtag). As base classifier we use either (1) a decision tree or (2) a logistic regression, both with balanced class weights. Since the performance of a decision tree strongly depends on its maximum depth and minimum leaf size, we perform grid search on the development set to select the optimal values for these parameters (see Table 5), using the same values for all base

#	Parameter name	Values
<i>Decision tree</i>		
1	maximum depth	5, 10, 15, 20, 25, ∞
2	min samples leaf	1, 2, 5, 10, 15, 20
<i>Logistic regression</i>		
1a	solver	newton-cg (ng), lbfgs (ls), sag (sg), saga (sa)
1b	multi class	multinomial (m), ovr (o)
2	C	.1, .5, 1, 5, 10

Table 5: Values for hyperparameters optimised in the grid search. The parameter number (#) and abbreviations in parenthesis are used in Table 6.

Setting	Multi									Binary								
	Development					Test				Development					Test			
	#1	#2	P	R	F	P	R	F		#1	#2	P	R	F	P	R	F	
<i>Decision tree</i>																		
split 1	5	20	.17	.72	.28	.28	.64	.39	25	1	.37	.54	.44	.45	.52	.49		
split 2	5	2	.13	.73	.22	.27	.61	.38	∞	10	.35	.69	.46	.43	.54	.48		
split 3	10	2	.17	.63	.27	.27	.58	.36	10	10	.38	.70	.49	.48	.63	.55		
majority	-	-	-	-	-	.28	.62	.38	-	-	-	-	-	.50	.60	.54		
<i>Logistic regression</i>																		
split 1	sg/o	.1	.19	.49	.27	.30	.51	.37	ls/m	.1	.40	.57	.47	.53	.64	.58		
split 2	sg/o	.1	.14	.57	.22	.29	.53	.37	sa/m	10	.37	.65	.47	.46	.57	.51		
split 3	ng/o	.1	.20	.53	.28	.31	.49	.37	ng/o	.1	.44	.64	.52	.55	.65	.60		
majority	-	-	-	-	-	.31	.51	.37	-	-	-	-	-	.54	.65	.59		

Table 6: Macro-averaged Precision (P), Recall (R) and F-score (F) on the development sets and the test set, for the multi-output and the binary classifier in all settings. Parameter values (#1 and #2, see Table 5) are given as optimised on the development set.

classifiers. For the logistic regression, we optimise the solver and multi-class parameter, and the regularisation parameter C .¹⁶ During the grid search, we use (macro-averaged) F-score (cf. Sokolova and Lapalme 2009) as scoring function, which we also use for evaluation.

Additionally, we combine the classifiers from the three splits into one majority classifier. The majority classifier assigns those tags to a clause that are predicted by at least two of the incorporated classifiers.

5.3 Evaluation

Table 6 shows Precision, Recall and F-score for all settings. For the binary classifier, Precision measures how many of the clauses tagged as comment are also annotated as comment in the gold standard; Recall measures how many of the clauses annotated as comment in the gold standard are also tagged as comment. The F-score is the harmonic mean of Precision and Recall. For the multi-output classifier, Precision, Recall and F-score are calculated separately for each subtag first and then averaged.

Decision tree and logistic regression show similar results on both the development sets

¹⁶ We set the maximum iterations of the logistic regression to 500. If not stated otherwise, we use scikit-learn's (<https://scikit-learn.org/stable/>) default parameters for our classifiers.

Setting	EINSTELLUNG		INTERPRETATION		META	
	Development	Test	Development	Test	Development	Test
	<i>#Clauses</i>					
split 1	173	201	154	252	172	280
split 2	171	--	216	--	73	--
split 3	258	--	147	--	31	--
	<i>Decision tree</i>					
split 1	.24	.28	.25	.38	.34	.50
split 2	.24	.29	.29	.31	.14	.52
split 3	.39	.26	.25	.34	.16	.50
majority	–	.28	–	.35	–	.52
	<i>Logistic regression</i>					
split 1	.26	.30	.26	.36	.31	.45
split 2	.24	.28	.29	.35	.13	.46
split 3	.39	.30	.28	.36	.17	.45
majority	–	.28	–	.36	–	.48

Table 7: Number of clauses and F-scores for each subtag on the development sets and the test set, for the multi-output classifier in all settings.

and the test set. The performance of both methods varies across splits, but the majority classifiers alleviate these discrepancies: In all but one setting, the majority classifiers achieve equal or better F-scores than the best of its incorporated classifiers.

Although the F-scores for decision tree and logistic regression are similar, Precision and Recall are not: The decision-tree classifier performs much better in terms of Recall at the cost of a lower Precision, whereas the difference between Precision and Recall is less extreme for the logistic-regression classifier.

Unsurprisingly, both methods achieve higher performance in the binary setting (54% and 59% for the majority classifiers) than in the multi-output setting (38% and 37% for the majority classifiers), where the classifiers have to distinguish subtags of comment.

5.4 Analysis

Somewhat surprisingly, every classifier performs better on the test set than on its development set. Part of an explanation might be that the test set includes more comment clauses than the development sets, see [Table 7](#), and our classifiers are mainly driven by Recall. [Table 7](#) also shows further differences between decision tree and logistic regression: With a logistic regression, the F-scores on the test set for each subtag are comparatively stable across training/development splits, whereas the decision tree's F-scores show a greater variance. The majority classifiers achieve performance close to the best individual classifiers for each subtag, resulting in F-scores of 28% for EINSTELLUNG, 35%–36% for INTERPRETATION and 48%–52% for META.

The comparatively high performance for META is outstanding, considering that META is the less frequent comment type in our data. In [Table 8](#), we calculate for each subtag the average number of training clauses that contribute to one percentage point on the test set. We can see that the ratio is significantly lower for META (12) than for EINSTELLUNG (103), with INTERPRETATION in between the two (60), something that illustrates that META

	EINSTELLUNG	INTERPRETATION	META
#Clauses	1887	2154	588
F-score	.28	.36	.48
Ratio (#/%)	103	60	12

Table 8: Number of clauses in the training set (including the development texts) and F-score of the logistic-regression majority classifier on the test set for each subtag. The bottom row shows the number of training clauses needed for one percentage point of F-score.

is much easier to learn by our classifiers than the other comment types.

Our binary classifier is considerably better than the multi-output classifier. In general, it is not unusual that a classifier performs better for a binary tagset than a more differentiated one. Still, since we observed in the agreement that annotators tend to disagree between EINSTELLUNG and INTERPRETATION while agreeing that a passage is one of both (see section 4), we trained an additional logistic-regression majority classifier that regards EINSTELLUNG and INTERPRETATION as the same tag. (We left out all META passages for this.) This classifier achieves an F-score of 47% on the test set, which is 19% higher than that for EINSTELLUNG and 11% higher than that for INTERPRETATION. Therefore, we assume that the difficulty of differentiating between EINSTELLUNG and INTERPRETATION applies for both humans and machine-learning methods, whereas a joint category is easier to learn.

#	EINSTELLUNG			INTERPRETATION			META			
	\pm	<i>i</i>	Unit	Feature	Value	\pm	<i>i</i>	Unit	Feature	Value
1	+	o	verb	mood	subj:past	+	o	verb	mood	subj:past
2	+	o	clause	speech	direct	-	o	clause	speech	direct
3	-	-1	clause	punct:inner	:	-	o	NP:obl	pos	PROPN
4	+	o	NP:nsubj	quant:type	NEG	-	-1	NP:obl	pos	PROPN
5	-	o	NP:obl	pos	PROPN	-	o	NP:nsubj	person	1per
6	+	1	verb	mood	subj:past	+	-1	verb	mood	subj:past
7	+	o	NP:root	emotion	Trust	+	1	verb	mood	subj:past
8	-	o	verb	mood	subj:pres	-	1	NP:obj	quant:pos	PRON
9	-	-1	NP:nmod	case	acc	-	1	clause	punct:prec	»
10	+	o	NP:advmod	art:pos	DET	-	o	NP:nsubj	person	2per
11	-	1	NP:obj	quant:type	DIV	-	-1	NP:nsubj	person	2per
12	+	o	clause	punct:succ	!	-	o	NP:obj	person	1per
13	+	1	NP:root	adj:category	Gefuehl	-	o	verb	mood	imp
14	-	-1	clause	punct:prec	:	+	o	verb	dep	csubj
15	+	1	clause	pos	X	-	1	NP:nsubj	person	1per
16	+	o	NP:root	emotion	Joy	+	o	verb	modal	scheinen
17	-	o	verb	tense	past	+	o	NP:nmod	art:lemma	ein
18	-	o	NP:obj	quant:type	DIV	+	1	verb	modal	scheinen
19	+	-1	NP:ccomp	numerus	sing	-	o	NP:appos	gender	masc
20	-	1	verb	mood	subj:pres	-	o	clause	punct:prec	«

Table 9: Top-20 features for each subtag ranked by absolute value of feature coefficient in logistic regression (split 1). \pm is the sign of the coefficient. *i* denotes whether the feature is extracted from the preceding (-1), current (o) or succeeding (1) clause.

Table 9 exemplarily shows the most important features for one logistic-regression classifier.¹⁷ Positive features are indicative for a subtag whereas negative features are indicative against a subtag.

Tense and mood/modality are learned to be relevant for all subtags. We have seen this in (3), where tense and mood shift from present indicative to past subjunctive to express a

17. The most important features show only minor variations between splits.

comment of type EINSTELLUNG. From the table, we can conclude that all three types often occur in past subjunctive, accompanied by different modal verbs (e.g. *scheinen* ‘seem’, *lassen* ‘let’, *wollen* ‘want’).

The comment types also differ in their presence within direct speech. While comments of type EINSTELLUNG frequently occur in direct speech, comments of type META rather occur outside direct speech. An explanation for this might be that utterances of characters in direct speech qualify for EINSTELLUNG, whereas META is mostly produced by the narrator. For INTERPRETATION, the speech feature is not important. Instead, it is learned that comments of type INTERPRETATION do rarely occur after quotation marks (« and »), which makes sense because they indicate a change of the speaker (from narrator to character or vice versa) and an interpretative comment typically follows a statement by the same speaker.

As anticipated in subsection 5.1, Example (5), a striking characteristic for EINSTELLUNG is the high importance of features related to emotion (Trust, Joy, and the more general feature *Gefuehl* ‘emotion’). For INTERPRETATION, we find that a subject in first person (*I, we*) or second person (*you*) is a negative indicator since only in third-person sentences something is told/interpreted about persons, incidents etc. For META comments, past tense is a negative feature. Instead, they often occur in grammatical future tense or with the modal verb *wollen* ‘want’, which can also express semantic future. This is illustrated in (6), where we can also see the typical use of *Kommunikation* ‘communication’ verbs, such as *erzählen* ‘tell’.

(6) Ein Märchen will ich dir erzählen_{META}, horche wohl. (Novalis 2012 [1802])

In general, our classifiers tend to return many shorter comment passages, with interruptions between them, while we annotate longer passages in the gold standard. This is because we train the classifiers on the clause level, giving only three clauses as input, whereas human annotators can draw connections between clauses that are farther apart. We do not see this as a problem, as long as the relevant passages from a text are returned. (Example 7) compares the gold annotations (a) and the predictions by the logistic-regression majority classifier (b) for an excerpt from Wieland’s *Geschichte des Agathon*. For sake of illustration, EINSTELLUNG is **boldfaced**, INTERPRETATION is *italicised* and META is underlined.

(7a) [...] so war es um so viel nötiger ihm auch dieser Probe zu unterwerfen, da Hippias, bekannter maßen, eine historische Person ist, und mit den übrigen Sophisten derselben Zeit sehr vieles zur Verderbnis der Sitten unter den Griechen beigetragen hat. [...]

(7b) [...] so war es um so viel nötiger ihm auch dieser Probe zu unterwerfen, da Hippias, **bekannter maßen, eine historische Person ist, und mit den übrigen Sophisten derselben Zeit sehr vieles zur Verderbnis der Sitten unter den Griechen beigetragen hat. [...]**

The excerpt is part of a long META passage in which the narrator reveals the conception of the main character Agathon and his confrontation with the sophist Hippias. The narrator outlines parts of the story, which can be seen as background knowledge that qualifies for an (overlapping) INTERPRETATION passage. Both passages span several (≥ 8) sentences in

#	±	<i>i</i>	Unit	Feature	Value	Subtags (#)
1	+	0	verb	mood	subj:past	EINSTELLUNG (1), INTERPRETATION (1), META (8)
2	-	0	NP:obl	pos	PROPN	INTERPRETATION (3), EINSTELLUNG (5)
3	-	0	NP:nsubj	person	2per	INTERPRETATION (10)
4	-	1	NP:appos	case	nom	
5	-	-1	NP:obl	pos	PROPN	INTERPRETATION (4)
6	+	0	verb	dep	csubj	INTERPRETATION (14)
7	-	0	clause	punct:prec	«	INTERPRETATION (20)
8	+	1	verb	dep	csubj	
9	+	-1	NP:nsubj	emotion	Fear	META (18)*
10	-	0	clause	dep	flat	
11	-	0	verb	tense	past	META (1)
12	+	0	clause	speech	direct	EINSTELLUNG (2), META (2)*
13	+	1	clause	punct:inner	«	
14	+	0	verb	modal	scheinen	INTERPRETATION (16)
15	-	0	verb	mood	imp	INTERPRETATION (13)
16	+	1	verb	mood	subj:past	META (4), EINSTELLUNG (6), INTERPRETATION (7)
17	-	-1	NP:appos	case	nom	
18	-	-1	NP:nsubj	person	2per	INTERPRETATION (11)
19	+	0	NP:obl	quant:pos	PART	
20	+	-1	NP:nmod	art:lemma	dies	

Table 10: Top-20 features for the binary classification ranked by absolute value of feature coefficient in logistic regression (split 1). ± is the sign of the coefficient. *i* denotes whether the feature is extracted from the preceding (-1), current (0) or succeeding (1) clause. The last column shows the rank of the features if it appears among the most important features for the individual subtags in Table 9. A star (*) indicates that the feature has the opposite sign in the subtag's base classifier.

the gold standard. The classifier detects shorter passages instead. It correctly recognises large parts of the excerpt as META. This is remarkable since the comprehension of the META passage is tied to its long context and our clause-based classifier is able to detect important parts of it. It also identifies large parts as INTERPRETATION, but is missing the beginning and a short interruption. Lastly, it also identifies the EINSTELLUNG in the last part of the excerpt; as well as a short EINSTELLUNG passage which is not in the gold standard. The short passage is a good example for a false positive: It is probably labelled as EINSTELLUNG because it features the evaluative term *bekanntes Maß* 'as is well known', but it does not express attitude towards the diegesis and is therefore not an EINSTELLUNG comment in the gold standard.

Table 10 shows the most important features of the binary classifier. It mostly includes important features for INTERPRETATION (see Table 9), which is the most frequent class in the training data. It also includes some important features for EINSTELLUNG, but important features for META are underrepresented, and there are even features with an opposite sign to those for META. This suggests that META passages are not individually learned by the binary classifier. This is not surprising when looking at Figure 1: Only 8% of all clauses in the training data are only annotated with META (other META clauses overlap with another comment type).

6. Discussion

6.1 General Considerations

As announced in the introduction, we do not consider the attempt to recognise literary comments as an end in itself. Rather, we want to use this example to illustrate that attempts to operationalise and recognise literary phenomena automatically can shed

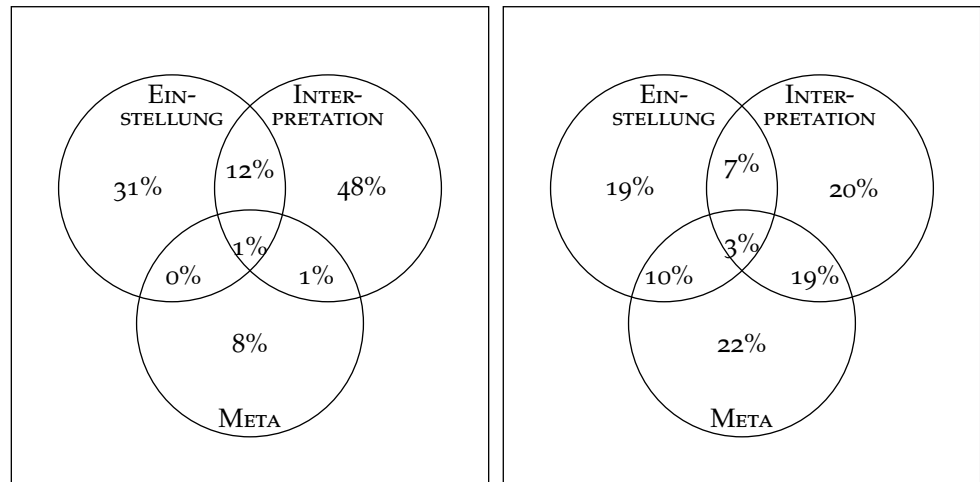


Figure 1: Overlap of comment clauses in the training data (left) and the test data (right).

light on the consistency of the concepts on which they are based.

When speaking of the ‘consistency of a concept’, we mean that (a) comparatively homogeneous phenomena fall under it and (b) that the concept is methodologically guiding in the sense that these phenomena are intersubjectively and automatically recognisable. Inconsistent concepts, on the other hand, describe comparatively heterogeneous phenomena, which cause hardly or not at all surmountable difficulties when trying to recognise them intersubjectively and/or automatically. Accordingly, ‘(in)consistency’ is a gradual concept: a concept can be more or less (in)consistent as the phenomena that fall under it have more or less relevant commonalities.

In the following, when we try to judge whether there is (in)consistency of a theoretical concept based on our observations on theory, operationalisation, annotation, and detection of it, we implicitly or explicitly carry out inferences to the best explanation. Generally, inferences to the best explanation have the following structure (see Lipton 2005; Bartelborth 2017, 200–291; here according to Descher and Petraschka 2019, 75):

P_1 : X is a fact that requires explanation.

P_2 : The hypothesis H_1 explains X .

P_3 : No competing hypotheses H_2, H_3, \dots, H_n explain X better than H_1 .

C: So H_1 is probably true.

Due to premise P_3 , inferences to the best explanation are not truth-preserving, i.e., a true conclusion does not always follow from true premises. Even if one considers as many relevant alternative hypotheses as practically possible, one may simply miss a hypothesis that explains X better than H_1 . Thus, claims about the consistency or inconsistency of certain concepts based on results of annotation and automation should be understood as hypotheses to be tested by further research.

In the case of comment, two main facts seem to need explaining:

- X_1 : While two subtypes of comment (attitude comment and interpretative comment) can be annotated with little intersubjective agreement and detected with little success automatically, the opposite is true for meta comment.

- X_2 : Automatic detection of comment by the binary tagger works well, although there is (to our knowledge) no robust overarching literary definition of comment and 2 of the 3 comment-subtypes are poorly recognised.

In the following, we first discuss the facts X_1 , then X_2 .

6.2 The (In-)Consistency of each Comment Type

One of our most intriguing findings is that annotation and automatic detection struggle with the same types of comment, although the annotation is based on a passage's reading whereas the automatic detection is based on a passage's form. Interpretative comment and attitude comment were annotated with only moderate and fair agreement and their detection also performed poorly with F-Scores below 40% and 30%. In contrast, meta comment achieves substantial agreement and can be detected well, with F-Scores close to 50%. Taking into account the relation between available training data for each comment-type and the performance of the multi-output-classifier, showed that meta comment is much easier to learn than the other comment-types (see Table 8; META performing better than INTERPRETATION/EINSTELLUNG by a factor of 5 and 8.5). Do these surprising results suggest an inconsistency of the concepts 'interpretive comment' and 'attitude comment' as we operationalised them based on several theories of literary comment?

With respect to annotation, we suspect that mainly semantic properties (the occurrence of terms such as 'narrative', 'truth' or 'invented') of the meta comment passages are responsible for their good agreement. The vague terms 'attitude' and 'interpretation', on the other hand, made the development of precise annotation guidelines difficult. What was contentious in the discussion of concrete annotations was not only at what point something is an attitude or interpretation, but also where the difference between the two lies. For clarification, let us recall example (3) from Fontane's *Der Stechlin*:

- (8) "Wir glauben doch alle mehr oder weniger an eine Auferstehung" (das heißt, er persönlich glaubte eigentlich nicht daran), "und wenn ich dann oben ankomme mit einer rechts und einer links, so ist das doch immer eine genierliche Sache." (Fontane 2012 [1898])

In this direct discourse passage the attitude of the main character Dubslav von Stechlin to a second marriage becomes clear. As an argument he uses the assertion that everyone more or less believes in the resurrection and the bad reputation of appearing with two wives. In the brackets between the direct speech, however, the narrator formulates a second attitude of Dubslav: he does not believe in the resurrection. Since this statement is not made by Dubslav but the narrator, we annotate it as INTERPRETATION. Chatman attempts to distinguish between these two types on the basis of the judgment/evaluation-criterion. However, he leaves open at what point a statement is evaluative enough to be considered an evaluative 'judgment' rather than an interpretation. This turned out to be problematic. For example, is the use of a term like 'eagerly' sufficient to show that the speaker has a positive or negative attitude towards someone? We found that annotators, based on their reading impressions, answer such questions differently.

For automatic recognition, it is, among other things, the difference between interpretative comment and attitude comment that causes problems. If we train a classifier that treats

EINSTELLUNG and INTERPRETATION as one tag (binary classification without META), we obtain F-Scores that almost approximate the binary score (EINSTELLUNG + INTERPRETATION + META). A problematic indicator of the attitude in both annotation and automatic recognition are “thick concepts” such as *eagerly* or *miraculous*, which “combine evaluation and non-evaluative description” (Väyrynen 2021).

If we exclude obvious alternative hypotheses such as unqualified annotators, inadequate machine-learning models, or errors in the statistical analysis of our classifiers,¹⁸ our findings suggest that (in contrast to meta comment) interpretive comment and attitude comment, as we have operationalised them, are *not* consistent concepts. The phenomena that fall under these two concepts are evidently too heterogeneous to be reliably recognised by humans and computers. Our findings on the automation side also suggest that there may be a consistent concept that encompasses all phenomena that fall under ‘attitude comment’ or ‘interpretive comment’. Defining this concept conclusively, without resorting (exclusively) to the vague terms ‘attitude’ and ‘interpretation’, would be a future task for Literary Studies.

6.3 The Inconsistency of the Generic Concept ‘Comment’

Although literary theory does not provide a consistent definition of the overarching concept of comment, our binary classifier (differentiating between comment and non-comment) achieves good results (F-Scores close to 60%). On the one hand, this is not very surprising because the binary classifier has (a) more training data per category than the multi-label classifier and (b) binary categorisation is less demanding. On the other hand, the classifier seems to accomplish the very thing that literary theory cannot provide (yet): a possibility to identify comment as a general phenomenon. What does this mean for a narratological concept of ‘comment’? We have already noted in section 3 that comment as a literary phenomenon is sometimes defined *ex negativo*. Therefore, many researchers refrain from defining comment and take an additive approach: Thus, ‘comment’ is understood as a set of related phenomena (phenomenon1 OR phenomenon2 OR ...) whose commonalities are rarely discussed.

Our own approach takes a related route, by identifying three comment types that share narrative pause as common feature or prerequisite. Our proposal, having the following logical form: necessary feature AND (feature1 OR feature2 OR feature3), takes the form of what Fricke calls a “flexible definition” (Fricke 1981). However, we have seen that there is reason to believe that two of the criteria that our operationalisation of ‘comment’ uses (‘interpretative passage’, ‘attitude passage’ are themselves not consistent concepts (see section subsection 6.2). Thus, the question arises whether the generic concept ‘comment’ is a meaningful consistent literary category at all.

It is important to see that automatic detectability is no reliable indicator that there is an underlying consistent concept. Not everything computers can automatically recognise is based on a consistent concept. Suppose we define the concept ‘tapple’ as ‘being an apple or a table’. This would be a very inconsistent concept because the phenomena that

18. We exclude these alternative hypotheses as improbable on the basis that (i) our annotators have a sound background in German Philology and have considerable experience with annotating works of literature, (ii) employ comprehensive machine learning models, extracting a wide variety of features which range from structural to sentiment features and (iii) employ a well-tested machine learning suite.

fall under it have little in common except that they are material objects. Nevertheless, one could undoubtedly build a supervised model that recognises ‘tapples’; it would most probably use the features of apples on the one hand and the features of tables on the other. Please note, that this only *prima facie* contradicts what has been said on inconsistent concepts above. The difficulty with automatic recognition would be that the model would be highly susceptible to bias due to unbalanced training data: If the majority of the training instances are tables, apples will probably not be detected at all, because they share no relevant commonalities with tables.¹⁹

So how does our binary classifier work? Our comparison of the most prominent features between the binary classifier (comment vs. non-comment) and the multi-label classifier (EINSTELLUNG, INTERPRETATION, META) yields an interesting result. 13 of the 20 most prominent features are features that also play a role for the recognition of the comment types (see Table 10). More importantly, only two of these 13 features are among the 20 most prominent features of all three comment types (subjunctive in the current or succeeding clause) and 9 features are indicative of one comment type only (according to the multi-label classifier). If the classifier had learned a general concept of comment, one would expect two kinds of features to dominate: features that are indicative of all three comment types and/or completely new features that played no role for the multi-label classifier. Therefore, our analysis suggests that the binary classifier, at least partly, uses feature combinations that are indicative of certain *types* of comment to recognise comment. The fact that 10 out of 20 most prominent features of the binary classifier are important features for interpretive comment (being the most common type in our training set, see Table 8), dovetails nicely with our expectation that a model that reflects a concept which is to a certain degree inconsistent is highly susceptible to bias due to unbalanced training data. Taken together, our results can be regarded as evidence for ‘comment’ being a rather inconsistent literary concept. The best explanation for the classifier not learning a general concept of comment is that the concept subsumes relatively heterogeneous phenomena that do not share enough relevant commonalities.

We have already underlined that our conclusions in the discussion section are ultimately hypotheses for which we have found some evidence, if we concede certain assumptions. There is one more background assumption that is relevant for our conclusion in this section of the discussion. Like many researchers in the Digital Humanities, we assume that literary phenomena manifest themselves at multiple levels (cf. Underwood 2019, 42), meaning that if there were a consistent narratological concept of ‘comment’, it would be reflected in linguistically available features. This assumption, rarely made explicit, may be more justified for an essentially textual phenomenon as comment than for phenomena that include relational properties. Let us suppose this background assumption is justified, so that our results show that ‘comment’ is a rather inconsistent concept. This would mean fundamentally re-examining the category of ‘comment’ and asking whether the important phenomena worthy of investigation that it describes cannot be grouped differently and/or partially subsumed under other concepts such as ‘authorial intrusion’ (Dawson 2016), ‘digression’ (Esselborn 1997–2003), ‘factual discourse’ / ‘serious speech acts in fictional works’ (Konrad 2017; Klauk 2015), ‘reflective passage’ (Gittel 2022), or *Sentenz* (‘aphorism’, Reuvekamp 1997–2003). At least this procedure seems appropriate

19. As every analogy, our analogy has its limits. In particular, comment types can overlap, because of their textual extent, but apples and tables as material objects do not.

to us, assuming that literary concepts should be also suitable for quantitative research nowadays.

7. Conclusion

Andrew Piper noted, that we “do not have a clear picture of how emerging quantitative methods speak to the questions that matter within the discipline of Literary Studies.” (Piper 2018, 10) The present paper addressed this issue by investigating the extent to which inferences about the consistency or inconsistency of textual literary concepts can be drawn from attempts at annotation and automation. Concretely, we operationalised the literary concept of ‘comment’ and phenomena associated with it: attitude passages, interpretative passages and meta passages. We annotated a corpus and trained classifiers for the automatic recognition of comment and its subphenomena. We were able to show that the concepts of the subphenomena vary in consistency. While meta comments are readily identifiable, clear overlaps emerge between interpretative and attitude comment. We also discussed the extent to which comment in and of itself can be understood as a consistent concept or as a catch-all for rather heterogeneous phenomena and found evidence in favor of the second assumption. We thus illustrated one way in which digital methods can contribute to humanities research in general and to a better understanding of ‘comment’ as a literary concept in particular. We not only examined an important literary phenomenon more closely and made it identifiable, we also addressed the question of why concepts such as the ‘literary comment’ are sometimes difficult to operationalise, investigating how far the success or failure of operationalisation and automation can help exploring their consistency.

8. Data Availability

Data can be found here: <https://doi.org/10.5281/zenodo.6467062>.

9. Software Availability

Software can be found here: <https://doi.org/10.5281/zenodo.6466328>.

10. Author Contributions

Anna Mareike Weimer: Data curation, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing

Florian Barth: Project administration, Data curation, Formal analysis, Resources, Software, Writing – original draft, Writing – review & editing

Tillmann Dönicke: Data curation, Formal analysis, Resources, Software, Validation, Writing – original draft, Writing – review & editing

Luisa Gödeke: Data curation, Methodology, Writing – original draft, Writing – review & editing

Hanna Varachkina: Data curation, Investigation, Resources, Writing – original draft, Writing – review & editing

Anke Holler: Funding acquisition, Supervision, Writing – review & editing

Caroline Sporleder: Funding acquisition, Methodology, Supervision, Writing – review & editing

Benjamin Gittel: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing

References

- Bartelborth, Thomas (2017). *Die erkenntnistheoretischen Grundlagen induktiven Schließens. Induktion, Falsifikation, Signifikanztests, kausales Schließen, Abduktion, HD-Bestätigung, Bayesianismus*. Universität Leipzig, Institut für Philosophie. <https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa-220168> (visited on 09/30/2022).
- Bonheim, Helmut (1975). "Theory of Narrative Modes". In: *Semiotica* 14 (4), 329–344. [10.1515/semi.1975.14.4.329](https://doi.org/10.1515/semi.1975.14.4.329).
- Booth, Wayne C. (1983). *The Rhetoric of Fiction*. University of Chicago Press.
- Chatman, Seymour Benjamin (1980). *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press.
- Dawson, Paul (2016). "From Digressions to Intrusions: Authorial Commentary in the Novel". In: *Studies in the Novel* 2 (48), 145–167. [10.1353/sdn.2016.0025](https://doi.org/10.1353/sdn.2016.0025).
- Descher, Stefan and Thomas Petraschka (2019). *Argumentieren in der Literaturwissenschaft. Eine Einführung*. Reclam.
- Dönicke, Tillmann (2020). "Clause-Level Tense, Mood, Voice and Modality Tagging for German". In: *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*. Association for Computational Linguistics, 1–17. [10.18653/v1/2020.tlt-1.1](https://doi.org/10.18653/v1/2020.tlt-1.1).
- (2021). "Delexicalised Multilingual Discourse Segmentation for DISRPT 2021 and Tense, Mood, Voice and Modality Tagging for 11 Languages". In: *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*. Association for Computational Linguistics, 33–45. <https://aclanthology.org/2021.disrpt-1.4> (visited on 09/30/2022).
- Dönicke, Tillmann, Luisa Gödeke, and Hanna Varachkina (2021). "Annotating Quantified Phenomena in Complex Sentence Structures Using the Example of Generalising Statements in Literary Texts". In: *17th Joint ACL-ISO Workshop on Interoperable Semantic*, 20–32. <https://aclanthology.org/2021.isa-1.3/> (visited on 09/30/2022).
- Ekman, Paul (1992). "An Argument for Basic Emotions". In: *Cognition & Emotion* 6 (3-4), 169–200. [10.1080/02699939208411068](https://doi.org/10.1080/02699939208411068).
- Esselborn, Hartmut (1997–2003). "Digression". In: *Reallexikon der deutschen Literaturwissenschaft*. Ed. by Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, and Klaus Weimar. Vol. 1. De Gruyter, 363–364.
- Fleiss, Joseph L (1971). "Measuring Nominal Scale Agreement Among Many Raters." In: *Psychological Bulletin* 76 (5), 378.

- Fludernik, Monika (2003). "Metanarrative and Metafictional Commentary: From Metadiscursivity to Metanarration and Metafiction". In: *Poetica* 35 (1/2), 1–39. <https://www.jstor.org/stable/43028318> (visited on 09/30/2022).
- Fontane, Theodor (1995). *The Stechlin*, translated by William L. Zwiebel. Camden House. <https://books.google.de/books?id=7yh0ZJjNgxoC> (visited on 09/30/2022).
- (2012 [1898]). "Der Stechlin". In: *TextGrid Repository*. Digitale Bibliothek. <https://hdl.handle.net/11858/00-1734-0000-0002-AECD-2> (visited on 09/30/2022).
- Fricke, Harald (1981). *Norm und Abweichung. Eine Philosophie der Literatur*. Beck.
- Genette, Gérard (1994 [1972]). "Diskurs der Erzählung [Discours du récit]". In: *Die Erzählung*. Trans. by Andreas Knop. Fink, 9–191.
- Gittel, Benjamin (2022). "Reflexive Passagen in fiktionaler Literatur. Überlegungen zu ihrer Identifikation und Funktion am Beispiel von Wielands 'Geschichte des Agathon' und Goethes 'Wahlverwandtschaften'". In: *Euphorion* 116 (2), 175–191. <https://euph.winter-verlag.de/article/EUPH/2022/2/5> (visited on 11/12/2022).
- Gius, Evelyn and Janina Jacke (2015). "Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse". In: *Grenzen und Möglichkeiten der Digital Humanities*. Ed. by Constanze Baum and Thomas Stäcker. Sonderband der Zeitschrift für digitale Geisteswissenschaften. 10.17175/sb001_006.
- (2017). "The Hermeneutic Profit of Annotation. On Preventing and Fostering Disagreement in Literary Analysis". In: *International Journal of Humanities and Arts Computing* 11 (2), 233–254. 10.3366/ijhac.2017.0194.
- Gödeke, Luisa, Florian Barth, Tillmann Dönicke, Anna Mareike Weimer, Hanna Varachkina, Benjamin Gittel, Anke Holler, and Caroline Sporleder (to appear). "Generalisierungen als literarisches Phänomen. Charakterisierung, Annotation und automatische Erkennung". In: *Zeitschrift für digitale Geisteswissenschaften*.
- Goethe, Johann Wolfgang von (2012 [1809]). "Die Wahlverwandtschaften". In: *TextGrid Repository*. Digitale Bibliothek. <https://hdl.handle.net/11858/00-1734-0000-0006-6A93-D> (visited on 09/30/2022).
- (19–?). *Elective Affinities : a Novel*. <https://archive.org/details/electiveaffinitio0goetuoft/page/68/mode/2up> (visited on 09/30/2022).
- Hamp, Birgit and Helmut Feldweg (1997). "Germanet - A Lexical-Semantic Net for German". In: *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. <https://aclanthology.org/W97-0802/> (visited on 09/30/2022).
- Herrmann, J. Berenike, Karina Van Dalen-Oskam, and Christof Schöch (2015). "Revisiting Style, a Key Concept in Literary Studies". In: *Journal of Literary Theory* 9 (1), 25–52. 10.1515/jlt-2015-0003.
- Hoffmann, E. T. A. (2012 [1816/17]). "Der Sandmann". In: *TextGrid Repository*. Digitale Bibliothek. <https://hdl.handle.net/11858/00-1734-0000-0003-6A92-6> (visited on 09/30/2022).
- (1885). *The Sand-Man*, translated by J.Y. Bealby. Charles Scribner's Sons.
- Hölderlin, Friedrich (2019). *Hyperion, or the Hermit in Greece*. Ed. by Howard Gaskill. Open Book Publishers. ISBN: 978-1-78374-655-2. <https://www.openbookpublishers.com/product/941> (visited on 09/30/2022).
- (2012 [1797]). "Hyperion oder der Eremit in Griechenland". In: *TextGrid Repository*. Digitale Bibliothek. <https://hdl.handle.net/11858/00-1734-0000-0003-7CC8-A> (visited on 09/30/2022).

- Hundsnurscher, Franz and Jochen Splett (1982). *Semantik der Adjektive des Deutschen. Analyse der semantischen Relationen*. Forschungsberichte des Landes Nordrhein-Westfalen. VS Verlag für Sozialwissenschaften.
- Jockers, Matthew Lee (2013). *Macroanalysis. Digital Methods and Literary History*. University of Illinois Press.
- Klauk, Tobias (2015). "Serious Speech Acts in Fictional Works". In: *Author and Narrator: Transdisciplinary Contributions to a Narratological Debate*. Ed. by Dorothee Birke and Tilmann Köppe. De Gruyter, 187–212. [10.1515/9783110348552.187](https://doi.org/10.1515/9783110348552.187).
- Konrad, Eva-Maria (2017). "Signpost of Factuality: On Genuine Assertions in Fictional Literature". In: *Art and Belief*. Ed. by Ema Sullivan-Bissett, Helen Bradley, and Paul Noordhof. Oxford University Press, 42–62.
- Lahn, Silke and Jan Christoph Meister (2013). *Einführung in die Erzähltextanalyse*. Metzler.
- Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33 (1), 159–174. [10.2307/2529310](https://doi.org/10.2307/2529310).
- Levin, Beth (1995). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.
- Lipton, Peter (2005). *Inference to the Best Explanation*. Routledge (International library of philosophy).
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (2015). "The Unified and Holistic Method gamma (γ) for Inter-annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41 (3), 437–479. <https://aclanthology.org/J15-3003/> (visited on 09/30/2022).
- Mohammad, Saif and Peter Turney (2010). "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon". In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, 26–34. <https://aclanthology.org/W10-0204> (visited on 09/30/2022).
- (2013). "Crowdsourcing a Word–Emotion Association Lexicon". In: *Computational Intelligence* 3 (29), 436–465. [10.48550/arXiv.1308.6297](https://arxiv.org/abs/10.48550/arXiv.1308.6297).
- Novalis (1903). *The Disciples at Sais and Other Fragments*, translated by F.V.M.T. and U.C.B., introduction by Una Birch. Methen & Co. <https://archive.org/details/disciplesatsais00nova/> (visited on 09/30/2022).
- (2012 [1802]). "Die Lehrlinge zu Sais". In: *TextGrid Repository*. Digitale Bibliothek. <https://hdl.handle.net/11858/00-1734-0000-0004-6129-B> (visited on 09/30/2022).
- Nünning, Ansgar (2005). "On Metanarrative: Towards a Definition, a Typology and an Outline of the Functions of Metanarrative Commentary". In: *The Dynamics of Narrative Form*. Ed. by John Pier. De Gruyter, 11–58.
- Piper, Andrew (2018). *Enumerations: Data and Literary Study*. University of Chicago Press.
- Piper, Andrew, Sunyam Bagga, Laura Monteiro, Andrew Yang, Marie Labrosse, and Yu Lu Liu (2021). "Detecting Narrativity Across Long Time Scales". In: *CHR 2021: Computational Humanities Research Conference*. Vol. CEUR-WS 2989, 319–332. http://ceur-ws.org/Vol-2989/long_paper49.pdf (visited on 09/30/2022).
- Prince, Gerald (2003). "Commentary". In: *A Dictionary of Narratology*. University of Nebraska Press, 1980.
- Remus, Robert, Uwe Quasthoff, and Gerhard Heyer (2010). "SentiWS – a Publicly Available German-language Resource for Sentiment Analysis". In: *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, 1168–1171.

- Reuvekamp, Silvia (1997–2003). “Sentenz”. In: *Reallexikon der deutschen Literaturwissenschaft*. Ed. by Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, and Klaus Weimar. Vol. 3. De Gruyter, 425–427.
- Sokolova, Marina and Guy Lapalme (2009). “A Systematic Analysis of Performance Measures for Classification Tasks”. In: *Information Processing & Management* 45 (4), 427–437. [10.1016/j.ipm.2009.03.002](https://doi.org/10.1016/j.ipm.2009.03.002).
- Underwood, Ted (2016). “The Life Cycles of Genres”. In: *Journal of Cultural Analytics* 2 (2). [10.22148/16.005](https://doi.org/10.22148/16.005).
- (2019). *Distant Horizons. Digital Evidence and Literary Change*. University of Chicago Press.
- Vauth, Michael, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann (2021). “Automated Event Annotation in Literary Texts”. In: *CHR 2021: Computational Humanities Research Conference*. Vol. CEUR-WS 2989. http://ceur-ws.org/Vol-2989/short_paper18.pdf (visited on 09/30/2022).
- Väyrynen, Pekka (2021). “Thick Ethical Concepts”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts/> (visited on 09/30/2022).
- Willand, Marcus, Evelyn Gius, and Nils Reiter (2020). “SANTA: Idee und Durchführung”. In: *Reflektierte algorithmische Textanalyse*. De Gruyter, 391–422.
- Zeller, Rosmarie (1997). “Erzählerkommentar”. In: *Reallexikon der deutschen Literaturwissenschaft*. Vol. 1. De Gruyter, 505–506.

A. Appendix: Translations of Examples

- (1') She [Otilie] was introduced to the gentlemen, and was at once treated with especial courtesy as a visitor. Beauty is a welcome guest everywhere. (J. W. v. Goethe 19-?)
- (2') When Nathanael lay on the stone pavement with a shattered head, Coppelius had disappeared in the crush and confusion. Several years afterwards it was reported that, outside the door of a pretty country house in a remote district, Clara had been seen sitting hand in hand with a pleasant gentleman, while two bright boys were playing at her feet. From this it may be concluded that she eventually found that quiet domestic happiness which her cheerful, blithesome character required, and which Nathanael, with his tempest-tossed soul, could never have been able to give her. (E. Hoffmann 1885)
- (3') The next day, as they were walking to the same spot, Charlotte took the opportunity of bringing back the conversation to the subject, perhaps because she knew that there is no surer way of rooting out any plan or purpose than by often talking it over. (J. W. v. Goethe 19-?)
- (4') I'd happily promise this book the love of the Germans. But I fear some will read it like a compendium and be overly concerned with the fabula docet, whilst others will take it too lightly, and neither party will understand it. (Hölderlin 2019)
- (5') Whosoever wills to be well acquainted with her [the Nature's] Soul must seek her company with the Poet, for to him she is expansive and pours out her miraculous heart EINSTELLUNG. (Novalis 1903)
- (6') I will tell thee a tale META. Listen! (Novalis 1903)
- (7') [...] so it was all the more necessary to subject him also to this test, since Hippias, as is well known, is a historical person, and, with the other sophists of the same time, contributed very much to the corruption of morals among the Greeks. [...]
- (8') Happy days awaited him there, the happiest of his life. But they were of brief duration; the very next year his wife died. Taking another was not for him, in part because of a sense of order and in part for aesthetic considerations. "After all," he maintained, "we all believe more or less in a resurrection (which is to say he personally really did not), and if I put in an appearance up there with one woman on my right and another on my left, well, that's always sort of an embarrassing business." (T. Fontane 1995)