Formal Ontology in Information Systems
C. Trojahn et al. (Eds.)

doi:10.3233/FAIA241299

Interpreting Texts and Their Characters

Emilio M. SANFILIPPO [a,b,1], Claudio MASOLO [a],
Emanuele BOTTAZZI [a], and Roberta FERRARIO [a],
[a] *CNR ISTC Laboratory for Applied Ontology, Trento & Catania, Italy*
[b] *CESR University of Tours, Tours, France*

Abstract.
Research in Digital Humanities calls for computational systems to document, compare, and analyze interpretations of cultural artifacts such as literary texts. These systems are intended to support scholars, critics, and students by facilitating access to existing analyses of texts, identifying similarities and divergences between interpretations, and more. We propose an approach for documenting interpretations of literary characters, grounded in the empirical practices of literary interpretation to align closely with experts' methods. To achieve this, we remain neutral regarding the ontological status of characters, instead relying on formal approaches based on linguistics. We demonstrate how our approach can analyze relations between names of fictional characters across texts and authors, bridging discussions in analytic philosophy about identity with the interests of literary scholars.

Keywords. interpretation; literary characters; identity; texts; literature

1. Introduction

There is a long standing debate about *fictional entities* in literary studies and philosophy [1,2], these including Sherlock Holmes in Doyle's stories, Emma Bovary in Flaubert's novel, and many others. Ficta extend to any sort of thing in a literary text, as well as imaginary entities that children fantasize about in their games. Our focus will be limited to literary characters because of their relevance in literary studies, where it is still debated whether characters are pieces of writing, person-like entities or something else [3].

A literary scholar may be interested in analyzing the traits of characters with respect to some critical theories, or may look at the relationship between characters across multiple texts to analyze their similarities and departing points. A philosopher, especially in the tradition of analytic philosophy, may ask in which sense a character exists, if it exists at all, or what is the criterion for its identity. To make a long story short, literary and philosophical debates have been developing in parallel trajectories and with different attitudes [4]. In this fragmented picture, literary scholars are barely interested in philosophical discussions on the ontological existence of ficta or their metaphysical characterization, whereas they are strongly focused on how texts and characters are *interpreted*, on the basis of which theories, approaches, sources, etc.

From a computer science perspective, current efforts aim to develop systems to support scholars in their interpretation practices [5,6]. Ideally, a new generation of systems

[1]Corresponding Author: emilio.sanfilippo@cnr.it

is emerging not only to access simple data about texts, such as their provenance, but also to explore alternative ways in which scholars have interpreted texts across cultures and epochs. With the goal of supporting the documentation of interpretations of literary texts, we present here an analysis of the notion of interpretation grounded on *inference*, and then apply the analysis to some aspects relative to the interpretation of literary characters.

From a methodological stance, some clarifications are needed to frame our proposal. First, we assume that a literary text does not have a single prescribed meaning (content) but that it can be interpreted in various ways (for references, see [5]). As a consequence, we cannot simply document the way in which a text "depicts" a character, because a text must be received by someone to tell anything. The relationship between characters and interpretations is an important departure point with respect to the debate in analytic philosophy considering that mainstream theories on ficta do not ascribe any role to interpreters (see [2]). A philosophical theory emphasizing the connection between texts, characters, and interpretations is presented by Paganini [7]. She argues that a necessary condition for the existence of a fictional entity with respect to a text is that interpreters attribute a *single* content to the text, meaning that they adopt a unique interpretation. According to her view, a fictional text possesses a single content when interpreters, based on their interpretational dispositions (i.e., the interpretations they might possibly provide for a text), agree on a set of possible situations that adequately describe what the text conveys. While we agree with Paganini's idea of grounding ficta on interpretations, our scenario clashes not only with the idea that texts have unique single contents but also with the intuition that interpretations require to consider interpreters' *dispositions*. From a literary perspective, interpreters articulate their positions only to a certain degree of precision and completeness. Even if one acknowledges dispositions, it is not the case that all scholars' interpretations are necessarily made explicit in a debate.

Second, as we will see, we adopt an approach based on the idea that scholars express and communicate their interpretations through natural language statements, which – in scholars' views – *follow*, *can be inferred*, *from* the interpreted texts.

Third, and this is a fundamental point, our proposal remains "agnostic" concerning the ontological status of ficta, and it is compatible with both realist and anti-realist philosophical positions regarding their existence. Accordingly, we solely consider statements and potential agreements among interpreters formed upon them without detailing ontological considerations. As radical as this move could seem, it is legitimate with respect to the pragmatic dimension of literary interpretation where experts only seldom wear the "ontological microscope" to dispute in which sense, say, Emma Bovary exists.

In our proposal, one can compare the activity of interpretation that scholars pursuit to that of a *game* in the sense of Wittgenstein's *Philosophical Investigations* [8]. In this game, interpreters are the players, the statements they produce in their interpretations are their moves, and the texts they adhere to in their interpretations are the rules, i.e., the constraints to which they have to adhere for the interpretation of a text to be an interpretation of *that* text. In this sense our approach on interpretation is not only *empirical* but also *normative*. An important dimension of literary interpretative games is that concepts like winning or losing do not apply, since it is far more crucial to understand on what interpretations *converge*. Agreements among scholars occur as a linguistic fact: we grasp that their interpretations converge only if we presuppose that interpreters agree on the text and the additional judgments they can make based on it. Also, just as in the game of chess we have, for example, only access to the rules and moves and do not need to

share the same ontological assumptions about the nature of the pieces in the game, in interpretative games all we have is the text and its interpretations. To say that we converge on an interpretation, we do not need to have a specific ontology of chess pieces rather than another. To strengthen the similarity with games, we should not forget that it is possible to play chess blindly, that is, using only language; the wooden pieces and the chess board only simplify the game but they are not an integral part of it. Ontological matters about characters' existence are not pressing, also because in literature we do not always need to "unload" our stipulations. It is as if, while someone tells us a joke about a policeman, we ask the person telling it if they really know that policeman. There are cases where instead we are curious, where the narrative seems to have certain characteristics that make scholars research to see if it is possible to unload this question about the existence of that particular character being talked about. We are, in other words, willing to treat certain narratives as hypothetical, without compromising the meaning of the story or the characters. Some philosophers converge with the idea of literary theorists that in narrating a story there is something akin to mathematics in this sense; they both are, in a way, stipulative and exploratory: "let us see where and how far a given assumption or basic situation can lead us" [9]. According to some [4] this can be done without the need to resort to the concept of "fictional truth". In any case, it is not, we insist, a pressing problem at the level of interpretation that of referring to some sort of fictional reality.

The remaining of the paper is structured as follows. Sect. 2 introduces our proposal to represent interpretations through *commitments* expressed in natural language sentences. By adopting studies in linguistics, we show that the formal semantics of our approach is compatible with both realist and anti-realist positions on ficta. We apply our approach in Sect. 3 to analyze the relationship between names of fictional characters across texts and authors making a connection between philosophical discussion on ficta's identity criteria and similar sorts of considerations done in literary studies. Finally, Sect. 4 concludes the paper.

2. Interpretations of Texts

We propose that the interpretation of a text is a form of "extension" of the text, i.e., in a dynamic conception of meaning [10], an updating or explication of the information contained in the text. From this perspective, the aim of interpreting a text is to make hypotheses about its content according to a group of interpreters, rather than to determine its factual basis [11]. We therefore assume a sort of stipulative (or pretence) attitude of interpreters [4], wherein interpreting a text requires, above all, accepting what is written in the text even when it conflicts with prior knowledge. To identify such extensions of a text, we rely on interpreters' *commitments* to linguistic statements. When multiple interpreters share their commitments regarding a text, we obtain a shared interpretation of the text. That is, our attitude is also, in a way, empirical. Following Wittgenstein [8, sect. 242], we take into account the moves of the agents in the interpretative game by considering them as empirical sentences about the text to be interpreted: measuring (interpreting) is determined not only by sharing methods of measurement (committing to the same text), but also by constancy in measurement results (sharing of judgments regarding texts).

A *text* T is here understood as a sequence of *sentences* in a natural language.[2] We write $\texttt{COMMIT}(a,s,T)$ to represent that the interpreter a makes public that s can be inferred from T, i.e., a commits to the fact that s follows from T, where s and T are sentences of the same natural language, and a is a competent speaker of this language. In our scenario focused on the documentation of scholars' interpretations, it is important to have intersubjective access to others' positions. $\texttt{COMMIT}(a,s,T)$ must be then understood in a *public* and *communicative* perspective, i.e., as a sort of *speech-act* and, more specifically, an *assertive* speech-act [12] performed by a.[3]

The general idea is that $\texttt{COMMIT}(a,s,T)$, by using a linguistic modality, communicates the result of a's inference processes: reading the text T, interpreter a dynamically builds a given body of information from which, according to additional (implicit or explicit) information a relies on, a can infer the information provided by s according to a's personal reading of s. In other words, by accepting what is reported in T, a also accepts what is reported in s or, in terms of "extensions" of T, $T \circ s$ is an acceptable extension of T according to a (where $T \circ s$ stands for the sequence of sentences obtained by adding the sentence s to the sequence T). We will show that the approach (i) is compatible with different theories of meaning, and (ii) does not presuppose a specific (shared) ontology. In this view, it is therefore possible that $\texttt{COMMIT}(a,s,T)$ and $\texttt{COMMIT}(b,s,T)$, where a and b do not share any ontology, nor the approach requires a to access the way in which b semantically grounds s (and T), and vice versa.

To further clarify our notion of commitment, given the linguistic nature of T and s, we find it useful to consider approaches in categorical grammar within the formal semantics of natural languages. More specifically, we consider the *Discourse Representation Theory* (DRT) [13] and later extensions such as the *Segmented Discourse Representation Theory* (SDRT) [14] that widen Montague grammar to apply to sequences of sentences called *discourses*. In these approaches, discourses (which are close to our texts) are dynamically translated into *Discourse Representation Structures* (DRSs), which formally represent discourses. Following Montague grammar, this translation primarily relies on syntactic and grammatical bases. However, as discussed in detail in SDRT, DRSs can incorporate some lexical or common-sense knowledge, among other factors, assumed to be shared by all competent speakers of a language. Importantly, DRSs (and, indirectly, discourses) can be, in their turn, translated into first order (FO) formulas.

When a discourse is (syntactically and grammatically) ambiguous, and lexical or common-sense knowledge are insufficient to disambiguate it, different DRSs must be considered. The translation from discourses to DRSs is therefore a one-to-many relation, that is, the same discourse can be translated into alternative DRSs. In our practical scenario, one can think that interpreters can disambiguate texts on the basis of cognitive, cultural, psychological, etc. grounds. For this reason, we consider an *interpreter-dependent* translation from texts to FO-formulas where interpreters can be more selective than DRT: an interpreter does not necessarily solve all the ambiguities in a text, but they can select a subset of all the FO-formulas associated to the DRSs that translate the ambiguous text. Formally, $\tau(a,T)$ is the set of FO-formulas that, according to interpreter a, represents text T.[4] A complication arises when texts are (superficially) logically inconsistent, e.g.,

[2] For the sake of simplicity, we do not consider multi-language texts.

[3] We implicitly assume that for every interpreter a, $\texttt{COMMIT}(a,T,T)$ holds.

[4] When a recognizes the ambiguity of T, $\tau(a,T)$ contains the logical disjunctions of the formulas corresponding to all the DRSs associated to T accepted by a.

when a text explicitly claims something and its negation. One can think that the shared lexical and common-sense knowledge together with the dynamic interpretation of the text can solve these inconsistencies. Alternatively, one can try to manage logical inconsistencies, e.g., by means of paraconsistent logics [15], or by considering only consistent fragments of the obtained set of FO-formulas. As a simplification hypothesis, we assume that $\tau(a,T)$ is consistent, i.e., interpreter a is able to solve the possible inconsistencies.

As said, interpreters' commitments can be based on additional (personal or shared) knowledge. When such knowledge is inconsistent with $\tau(a,T)$, our notion of interpretation presupposes that only part of such knowledge can be used for their commitments. We indicate with $\kappa(a,T)$ the part of the knowledge of a that they select to solve the possible inconsistencies with $\tau(a,T)$. We assume that $\kappa(a,T)$ is also represented by means of FO-formulas. In Sect. 2.2 we will be more specific on the knowledge of interpreters by distinguishing lexical and common-sense knowledge from reference to other texts considered by a to interpret T.

At this point we can be more explicit about the requirements behind commitments: $\texttt{COMMIT}(a,s,T)$ requires that (i) $\tau(a,s)$ follows from $\tau(a,T)\cup\kappa(a,T)$ but (ii) $\tau(a,s)$ does not follow from $\kappa(a,T)$ alone, i.e., for a, what is written in T is necessary to commit to s.[5] Commitments are subjective to an interpreter a in two ways: (i) $\tau(a,s)$ and $\tau(a,T)$ depend on how a solves linguistic ambiguities and possible internal inconsistencies of s and T; (ii) $\kappa(a,T)$ depends on a's prior knowledge, as well as on the way in which a solves possible inconsistencies between their prior knowledge and $\tau(a,T)$.

We will show in the next sections how the notion of COMMIT is compatible with both realist and anti-realist positions on the literary characters featuring in T.

2.1. Commitments: Realist vs. Anti-Realist Positions

In a realist perspective, a semantic approach to inference can be embraced: proper names, definite descriptions, and indexicals are translated into (FO) individual constants which, in their turn, are mapped to elements of the domains of model-theoretic structures providing the truth conditions for formulas in $\tau(a,T)$, $\kappa(a,T)$, and $\tau(a,s)$.

In this perspective, we introduce a third element of subjectivity in commitments: interpreter a can have a specific *ontological view*, i.e., a can interpret the formal language in a restrictive way by considering a proper subset $\mathfrak{M}(a)$ of the whole set of the structures of the language.[6] The requirements on $\texttt{COMMIT}(a,s,T)$ can be then restated as: for any model $\mathscr{M}\in\mathfrak{M}(a)$, if $\mathscr{M}$ satisfies all the formulas in $\tau(a,T)\cup\kappa(a,T)$ (we write $\mathscr{M}\models\tau(a,T)\cup\kappa(a,T)$ for all $\phi\in\tau(a,T)\cup\kappa(a,T)$, $\mathscr{M}\models\phi$), then it also satisfies all the formulas in $\tau(a,s)$ (i.e., $\mathscr{M}\models\tau(a,s)$). Furthermore, there exists a $\mathscr{M}\in\mathfrak{M}(a)$ such that $\mathscr{M}\models\kappa(a,T)$ but $\mathscr{M}\not\models\tau(a,s)$.[7]

Some philosophers embrace anti-realist positions with respect to ficta (see [2]) where fictional names like 'Sherlock Holmes' do not refer to any entity. We will now

[5] By considering $\kappa(a,T)$ among the knowledge one can use to derive $\tau(a,s)$, we embrace a pretense mediated version of *Reality Assumption* [16], which has the known issue that everything which is in $\kappa(a,T)$ can be also the subject of a commitment. Even though clause (ii) mitigates the problem, still conjunctions of formulas in $\kappa(a,T)$ and in $\tau(a,T)$ could be included in $\tau(a,s)$. We do not consider this problem in the following.

[6] More specifically, one could assume that the ontological view of a is represented by means of simpler structures from which $\mathfrak{M}(a)$ can be (set-theoretically) constructed.

[7] Usually, $\tau(a,T)\cup\kappa(a,T)$ is not semantically complete, thus even when $\tau(a,T)\cup\kappa(a,T)=\tau(b,T)\cup\kappa(b,T)$, the structures in $\mathfrak{M}(a)$ and $\mathfrak{M}(b)$ are not necessarily isomorphic.

show that in this anti-realist view we can still embrace a semantic approach to inference by considering a psychologistic variant of the Tarskian definition of truth, where truth-conditions of fictional statements are grounded on interpreters' mental states.

Mental states can be represented by taking inspiration from the theory of *mental files* by Recanati [17]. In this line of works, Korta et al. [18] distinguish different kinds of statements, providing for them different sorts of referential or non-referential truth-conditions. We provide more details on the approach put forward by Maier [19], which is based on an extension of DRT allowing for a direct comparison with the previous reading of COMMIT. By relying on recent work by Kamp [20], Maier [19] extends standard DRT with mental attitudes, i.e., DRSs are paired with labels representing mental attitudes like *believing*, *desiring*, *intending*, etc. In particular, *imagining* is included among mental attitudes, due to Maier's reliance on Walton's approach [21], where fictional statements serve as prescriptions for imagination. In this psychologistic version of DRT, labeled-DRSs represent interpreters' mental states, which are dynamically updated during the interpretation of a discourse. Furthermore, following the idea of mental files, so-called *anchoring* mechanisms are introduced to indicate the "DRSs that serve as descriptive internal representations of objects the agent is acquainted with" [19, p.9].

Without entering into the details of Maier's approach [19], the crucial aspect is that the truth-conditions for fictional and non-fictional statements are provided in terms of how labeled-DRSs capture an agent's mental state. Labeled-DRSs are formally interpreted in terms of complex structures where, however, an external referent for the involved entities is not always presupposed (in particular for imagining). In his approach, Maier assumes a uniform semantics for both fictional and non-fictional statements – in particular, "Holmes lives in London" vs. "Holmes is a fictional character" – avoiding the problem of categorizing statements under sentence-kinds to which different truth-conditions apply. Note however that this approach is intrinsically based on the labeling of DRSs with mental attitudes, which – as admitted also by Kamp [20] – usually go beyond the scope of natural language semantics. This because the mental attitude that an interpreter adopts in response to a statement is not in the propositional content of the statement itself. The critical aspect linked to the identification of statement-kinds discussed in [18] seems therefore moved to the identification of interpreters' mental attitudes. One has in a sense a further dimension of subjectivity where the adoption of a mental attitude on the interpreter side relies on extra-textual information available to the interpreter; e.g., because the interpreter assumes to read a fictional text, her mental attitude is inclined towards imagination rather than belief. This overall picture does not impact our framework where commitments are assumed as being opaque to the specific mental attitudes of interpreters. Thus, the requirements on $\texttt{COMMIT}(a,s,T)$ above introduced can be maintained once the mental counterparts of $\kappa(a,T)$, $\tau(a,T)$, and $\tau(a,s)$ are considered.

2.2. Grounding Commitments on Additional Texts

Up to now, the knowledge an interpreter can use to make explicit some information in the text T, formally denoted as $\kappa(a,T)$, is a black box. One may assume that such knowledge includes some (minimal) lexical and common-sense knowledge shared by all competent speakers of a language, but in general, interpreters' knowledge can differ due to their experiences, readings, cultures, etc. In this section we will refine the notion of commitment to explicitly indicate when the information used to infer s originates

from other texts. By explicitly specifying the "sources" of the knowledge underlying a commitment, literary debates about the interpretations of texts can be better documented. For instance, to support the interpretation of a novel by Doyle, a scholar may use a text of criticism about Doyle. In a sense, with the support of these (critical) texts, interpreters explain literary texts *in the light of* other texts [22].

$\texttt{COMMIT}(a,s,T,U)$ stands for: "According to interpreter a, what is reported in sentence s follows from what is reported in text T, given what is reported in text U." The general idea is that an essential part of the information a adopts to commit to s derives, modulo linguistic disambiguation, from what is reported in U. Following the analysis in Sect. 2.1, $\texttt{COMMIT}(a,s,T,U)$ requires that $\tau(a,T) \cup \tau(a,U) \cup \kappa(a,T,U) \models \tau(a,s)$ but $\tau(a,T) \cup \kappa(a,T) \not\models \tau(a,s)$. Here $\kappa(a,T,U)$ represents the lexical and common-sense knowledge of a they select to solve possible inconsistencies with $\tau(a,T) \cup \tau(a,U)$, i.e., generalizing what done for $\texttt{COMMIT}(a,s,T)$, the interpreter a accepts what is reported in T and U, even though this goes against some common-sense knowledge. For instance, assume that (1) "Holmes lives in 211 Baker Street" and (2) "Baker Street is in London" are in T. If $\kappa(a,T)$ contains appropriate knowledge about the preposition "in", to commit to (3) "Holmes lives in London", a does not require additional information. On the other hand, if T and $\kappa(a,T)$ do not contain any information about the location of London, then (4) "Holmes lives in England" cannot be inferred from T. However, a can ground their commitment to (4) by referring to a text U containing (5) "London is in England."

Some simplification hypotheses shape our preliminary proposal. First, we assume a single supporting text U, but clearly a could need several texts to ground their commitment. Second, instead of grounding their commitment on what is reported in U, a could rely on some prior interpretations of U that, in their turn, can be supported by other texts, i.e., a chain of texts could be necessary in this case. We leave this extension for future work. Third, in the previous example, the sentence (2) is in T while the sentence (5) is in U. However, to derive (4) one needs to assume that the name London in T and the name London in U have the same meaning. For the moment we assume a default "same name / same meaning" attitude. However, there may be scenarios where identical names have different meanings or different names have the same meaning. Thus, $\texttt{COMMIT}(a,s,T,U)$ depends in general on some mappings between the proper names (or definite descriptions) appearing in T and U, see Sect. 3.

2.3. Commitment and Inference

Jacke [23] distinguishes *contextual* approaches to literature from *interpretative* ones. In contextual approaches, the understanding of a text can depend on additional available "material" (in Jacke's terminology). Differently, in purely interpretative approaches, the understanding of a text depends on the non-deductive nature of the adopted inferential mechanism allowing different subjects to produce, starting from the same material, different results. In Jacke's words, if "the inference rules allow for the derivation of different results from the same input material, and hence an individual has to decide which of the possible results is the correct one, a statement about meaning is interpretive" [23, p.130]. The subjectivity of texts' interpretation seems then based only on the presence of non-deductive inferential mechanisms.

In our framework, $\texttt{COMMIT}(a,s,T,U)$ allows us to explicitly represent the contextual – in the sense of Jacke – nature of the understanding of a text. However, our framework

is grounded on semantic approaches to reasoning where the notions of *truth-condition* and *truth-preservation* play a central role. One can then wonder whether COMMIT (and then interpretations, see Sect. 2.4) can be grounded on subjective forms of inference.

As already said, the translation of a linguistically ambiguous text into FO-formulas, and the solution of possible inconsistencies between the (translation of the) text and the knowledge of an interpreter both depend on the specific choices of the interpreter. Hence, even presupposing that all interpreters share the same input materials, still a subjective dimension can be present. Furthermore, interpreters can have different (mental) attitudes towards the interpreted texts grounded on different elements. Finally, interpreters can have different knowledge, and this knowledge is not necessarily made explicit via shareable material since it can be the result of complex learning processes.

The role of the ontological view of interpreter a, represented by $\mathfrak{M}(a)$ in Sect. 2.1, deserves a more detailed discussion. $\mathfrak{M}(a)$ is a subset of the whole set of the models of the formal language. In the translation of the requirements on COMMIT adopted in Sect. 2.1, we considered a form of (truth-preserving) deduction "localized" to $\mathfrak{M}(a)$, i.e., the truth is preserved inside the models preferred by a but not necessarily in general. A link with the semantic (model-theoretic) approaches to non-monotonic reasoning (see [24] for a recent systematic analysis), and in particular with the notion of *preferential entailment* introduced by Shoham in [25], can be established by refining the idea of $\mathfrak{M}(a)$. One can presuppose that the ontological view of interpreter a, rather than with $\mathfrak{M}(a)$, is represented by means of a (partial) order defined on the models of the formal language: $\mathscr{M} \sqsubset_a \mathscr{N}$ means that the model $\mathscr{M}$ is *preferred by a* over the model $\mathscr{N}$. One can then say that for the interpreter a, $\tau(a,T) \cup \kappa(a,T)$ *preferentially entails* $\tau(a,s)$ when for every $\mathscr{M}$ such that (i) $\mathscr{M} \vDash \tau(a,T) \cup \kappa(a,T)$ and (ii) there is no model $\mathscr{N}$ such that $\mathscr{N} \vDash \tau(a,T) \cup \kappa(a,T)$ with $\mathscr{N} \sqsubset_a \mathscr{M}$, then $\mathscr{M} \vDash \tau(a,s)$. Adopting this reading, COMMIT is grounded on a kind of preferential entailment that in general is nonmonotonic and where the "outputs" of the inference process subjectively depend on the preference relation $\sqsubset_a$ embraced by interpreter a.

Another form on non-monotonicity can emerge when considering $\mathtt{COMMIT}(a,s,T,U)$ because nothing guarantees that $\mathtt{COMMIT}(a,s,T) \rightarrow \forall U(\mathtt{COMMIT}(a,s,T,U))$. This is due to the fact that the relation between $\kappa(a,T,U)$ and $\kappa(a,T)$ is not constrained, i.e., a can resolve possible inconsistencies discarding different parts of their original knowledge.[8]

2.4. From Commitments to Agreements and Interpretations

Commitments can be easily generalized to the truth of a whole text S instead of a single sentence s; it is sufficient to consider $\tau(a,S)$ instead of $\tau(a,s)$ in the previously discussed requirements. One can then define the agreement of a set of agents A on a given interpretation of T given U as follows: $\mathtt{AGREE}(A,S,T,U) := \forall a \in A(\mathtt{COMMIT}(a,S,T,U))$.

The notions of commitment and agreement can be further generalized (noted with gCOMMIT and gAGREE) by allowing agreements based on commitments grounded on different sources of information, i.e., the interpreters in A can support their commitments taking into account different texts: $\mathtt{gCOMMIT}(a,S,T) := \exists U(\mathtt{COMMIT}(a,S,T,U))$, $\mathtt{gAGREE}(A,S,T) := \forall a \in A(\mathtt{gCOMMIT}(a,S,T))$.

An interpretation of a text T is a maximal text S on which a set A of agents agrees: $\mathtt{INT}(A,S,T)$ stands for "the text S is the interpretation of the text T from the point of view

[8] Note that DRT and SDRT often involve non-classical logics, e.g., dynamic and nonmonotonic logics.

of the set of agents A."[9] Thus, following Wittgenstein again, the understanding achievable through language does not depend only on the agreement we have on our rules (i.e. on the constraints imposed by the text to be interpreted), but also on the agreement we have regarding our moves (i.e. on the judgments we express in our interpretations). In this perspective, the issue of the ontological nature of the entities we talk about may become superfluous, i.e., the interpretation of a text may not be affected by the different positions regarding the nature of these entities, be they realist or anti-realist. Analogously, Hirsch [26] observes that discussants (in philosophical debates in metaphysics) do not necessarily need to share a common ontological view to understand each other. This is because each discussant can make sense of what others say based on their own ontological view and shared principles of conversation. But clearly it is possible to have different sets of interpreters agreeing on different interpretations of the same texts, e.g., $\texttt{INT}(A,R,T)$ and $\texttt{INT}(B,S,T)$ with $R \neq S$.[10]

For the sake of clarity, although we have followed Paganini's [7] idea of grounding a text's interpretation in the agreement on S among interpreters of the text T, our approach differs significantly from hers. Paganini envisions a sort of ideal case where all interpreters need to agree on what is included in the content of a text. This agreement is understood in terms of the *dispositions* that interpreters have to accept a statement. Differently, following the practices of literary interpretation, we allow interpreters agreeing on different (possibly inconsistent) and partial contents, where agreements are the result of a commitment on such partial contents. Furthermore, it is important to stress that $\texttt{AGREE}(A,S,T,U)$ does not exclude the possibility to have interpreters in A assuming different translations of T, U, and S, as well as different ontological commitments. Furthermore, $\texttt{gAGREE}(A,S,T)$ abstracts from the additional texts on which interpreters in A base their commitments. Interpreters might therefore have different reasons to commit to S. Even if we presuppose that all interpreters have a realist reading, possess the same common-sense knowledge, resolve linguistic ambiguities and logical inconsistencies in the same manner, and support their commitments with the same text U, $\texttt{AGREE}(A,S,T,U)$ does not imply that all interpreters in A share the same ontological view, since each interpreter can consider in $\mathfrak{M}(a)$ very peculiar models that no other interpreter in A considers. It should also be noted that the specific positions, translations, knowledge, and ontological commitments of agreeing interpreters are not generally public and accessible to other interpreters.

In the next section, we explore how our notions of commitment, agreement, and interpretation can provide an empirical basis to relate characters' names found in different texts and authors, as well as to distinguish between fictional and non-fictional names.

3. Relations between Characters' Names

As we have seen in Sect. 2.2, interpreters can approach multiple types of texts, ranging from literary texts to other sorts of texts used to support their interpretations. In all these cases, they need to understand whether the linguistic elements (names, descriptions, etc.)

[9] One may introduce different notions of interpretation grounded on different notions of agreement. We do not enter into these details here since they are not relevant for the following discussion.

[10] The relation between R and S can be better qualified by introducing shared notions of non-equivalence or incompatibility, or by introducing additional kinds of speech-acts like rejection, doubt, etc.

n and m in texts T and U, respectively, have the same meaning. Our idea is to address this question in the light of debates around identity, similarity or other types of relations between literary characters (see [2] for some discussions).

We consider a simplified scenario sufficient for illustrating how different positions regarding identity can be reconstructed within our approach. First, we focus solely on proper names, excluding definite descriptions and indexicals. Second, to determine whether two names appearing in different texts have similar (or identical) meanings, we rely only on the partial information provided in the texts (according to given interpretations) and additional agreements about the fact that certain sentences characterize the meaning of such names.[11] Third, while a name may have different meanings in different texts, we assume that a name maintains the same meaning within single texts.

3.1. Diagnostic Traits

Among all traits that scholars may consider to identify and analyze characters, they might focus on a subset, considered as particularly relevant. To report an example, discussing about the character of Emma Bovary, Eco [27] claims that the character in Gustave Flaubert's novel *Madame Bovary* and Woody Allen's film script for the *The Kugelmass Episode* is the same one, although in the latter case she does not commit suicide and behaves as a Tiffany-goer. This because Emma Bovary is still recognizable since "she keeps most of her basic properties – namely, she is a petty bourgeois and the wife of a doctor, she lives usually at Yonville, she is unsatisfied with the countryside life, she is inclined to adultery." Eco concludes that "a fictional character remains the same even if it is set in a different context, provided *diagnostic* properties (to be defined for each case) are preserved" – on similar lines, see also Richardson [28]. As a perhaps less controversial example, in the case of *series* like the stories of Doyle, intuitively Holmes remains the same character even though across the books of the series he might have different "marginal" traits and just maintains the relevant ones. Let us then assume that characters are associated with *diagnostic traits* used to identify them across texts. We will explore how this approach can be represented and exploited in our framework.

In our perspective, we take into account characters by considering their names and what is said about them. Interpreters may commit to multiple sentences involving a name n in T, while considering only some of them as salient for n. In other words, interpreters *select* the relevant sentences for n, corresponding to the diagnostic traits, from those in their interpretations of T.

To represent the selection of diagnostic traits, we introduce a new kind of commitment: $\texttt{TRAIT}(a,R,n,T)$ stands for "the interpreter a declares that the sentences in R are all relevant for the name n as appearing in the text T",[12] where $\texttt{TRAIT}(a,R,n,T) \rightarrow \texttt{COMMIT}(a,R,T)$, i.e., the relevant sentences for n in T are included in a's interpretation of T. Following what done for interpretation, $\texttt{TRAIT}(A,R,n,T)$ collects in R all relevant sentences for n in T on which the set A of interpreters agrees, i.e., we have that $\texttt{TRAIT}(A,R,n,T) \rightarrow \texttt{AGREE}(A,R,T)$.

In $\texttt{TRAIT}(A,R,n,T)$, R can be seen to correspond to diagnostic traits, but note that (i) R contains sentences relative to a *name* as it appears in a text, and (ii) R expresses

[11] For simplification purposes, we assume that these sentences characterize the meanings of a name only intrinsically, i.e., they do not concern relations with other names.

[12] Given our simplified scenario, R does not contain relational constraints between names.

only the point of view of the set A of interpreters. Said that, the criterion (C1) establishes when, for A, the names n and m have the same diagnostic traits – where $V_{(n_1 \to n_2)}$ indicates the text obtained by syntactically substituting in V the name n_1 with the name n_2 – by assuring that (i) A selected the diagnostic traits for both n in T and m in U; and (ii) modulo the lexical/common-sense knowledge, such traits are equivalent. (C1) can be weakened as in (C2). In this case, A selected the diagnostic traits only for n in T and A just recognizes that such diagnostic traits for n apply also to m in U. (C1) and (C2) can be further weakened to individuate partial matches between the traits associated to names, e.g., by assuming that (C1) and (C2) hold only for a proper subtext of R (and S).

C1 According to the set A of interpreters, name n in text T and name m in text U have the same diagnostic traits if and only if there exist texts R and S such that (a) $\texttt{TRAIT}(A,R,n,T)$ and $\texttt{TRAIT}(A,S,m,U)$; and (b) for all $a \in A$, $\kappa(a,T) \models \tau(a,R) \leftrightarrow \tau(a,S_{(m \to n)})$ and $\kappa(a,U) \models \tau(a,S) \leftrightarrow \tau(a,R_{(n \to m)})$.

C2 According to the set of interpreters A, name m in text U satisfies the diagnostic traits of n in text T if and only if (a) there exists a text R such that $\texttt{TRAIT}(A,R,n,T)$; and (b) $\texttt{AGREE}(A,R_{(n \to m)},U)$.

(C1) and (C2) (as well as their weaker versions) establish links between names as appearing in texts independently of any (historical) evidence concerning the relationships between such texts or their authors. That is, such identities and similarities could be just "fortuitous" or "unintentional". This could be a legitimate perspective when scholars may wish to study characters only by considering their traits independently from any other information. On the other hand, following philosophical positions like *artefactualism* on fictional entities [2], one may restrict the previous criteria to texts with the same author in order to study relations between characters by taking into account not only diagnostic traits but also authorship. Further refinements can be introduced when additional information is available, e.g., concerning the time at which texts were produced.

(C1) (or (C2)) can be assumed to be enough to conclude that n in T and m in U have the same meaning, i.e., they are interchangeable not only contextually to their diagnostic traits, but in all the sentences.[13] T and U offer then a sort of unified view on the character named n or, interchangeably, m, i.e., it is like having a single text (say $T \circ U$ composed by T and U) talking about a single character. However, $T \circ U$ collects all the traits of n in T and m in U, even when there are inconsistencies between them, i.e., $T \circ U$ could result a superficially inconsistent text. For instance, Eco seems to suggest that the character of Emma Bovary in Flaubert's work and Emma Bovary in Allen's work have the same diagnostic traits and can therefore be identified, even though in Flaubert's work she commits suicide while in Allen's work she does not. As in the case of a single inconsistent text, one can assume that the interpreters are able to solve these inconsistencies, for instance by finding some reasons to exclude one of the two contrasting traits from $T \circ U$. Alternatively, one may assume that when inconsistent traits are identified for a fictional name, different characters have to be distinguished. In a perspective aligning with philosophical possibilism, the two characters may stand in counterfactual modal relations, rather than identity, each one being characterized only through consistent traits.

From this latter perspective, one might assume that (C1) (or (C2)) does not imply identity of meaning but weaker relations, including what is sometimes called *borrowing*.

[13] This would be analogous to an identity criterion for characters based on diagnostic traits.

For instance, one may claim that an author has borrowed a character from another author, adapting him/her to their own specific narrative goals. In this case, one may think that diagnostic traits must be preserved when borrowing characters, i.e., borrowing is a case of identity-preservation. Clearly, the recognition of authors' intentions to borrow characters from other texts can be problematic, especially without accessible evidence in written sources on which scholars may rely. Borrowing can be in its turn weakened into *derivation* by considering weaker versions of (C1) (and (C2)), i.e., derivation may require only an (partial) overlap between the traits associated to the names that may not even capture diagnostic traits.[14] In the case of Emma Bovary in Flaubert and Allen, different positions can be therefore assumed depending on the manner in which diagnostic traits, relations between texts, and authors' intentions, among other dimensions, are evaluated. For instance, assuming that Allen wanted to establish a connection with Flaubert's literary work, if her death in Flaubert's novel is considered as a diagnostic trait, one may consider – *contra Eco* – Allen's Bovary as a derivative character sharing with Flaubert's Bovary several traits without for this being identical with it. Alternatively, one may tend to think her death as a marginal trait, in which case it could be claimed that Allen has borrowed Flaubert's character while recontextualizing her for his purposes.

In a nutshell, reference to diagnostic traits can be useful to support the analysis of relations between literary characters by looking at the interpretations related to their names in texts, even when the texts are not of the same author. In addition, one may consider further information about texts and authors when reference to diagnostic traits is not sufficient to ground meaningful literary relations between characters. From this perspective, it is important to stress that differently from mainstream philosophical theories on literary characters, the individuation of diagnostic traits or relations between texts and authors are always given in hypothetical terms, that is, they are considered from an empirical, scholarship perspective rather than being based on metaphysical principles.

3.2. Towards an Empirical Grounding for Ficta

The way in which texts are interpreted may help in supporting the distinction between fictional and non-fictional entities. The distinction is nuanced from a literary standpoint. For some characters like Holmes or Emma Bovary, scholars do not have doubts about their fictional status. For others the debate is more controversial. To make an example, the figure of Francesca da Rimini in Dante's *Comedy* recalls the story of Francesca da Polenta, a woman who lived during Dante's life. However, there remain only few historical traces about her so that scholars are cautious in identifying Dante's Francesca with Francesca da Polenta. One may ask whether Francesca da Rimini is a fictional character on the lines of Holmes and Emma Bovary; something similar could be asked for other characters, including Napoleon in Tolstoy's *War and Peace*, among others. To add more complexity, there are several figures for which scholars do not know whether they are fictional or historical, like in the case of Boccaccio's text *De mulieribus claris*.

In our approach the boundary between "reality" and "fiction" can be traced based on the relationship between texts (see [28] on similar lines). The boundary, in other words, is drawn by relating, for example, Napoleon in *War and Peace* to Napoleon in an encyclopedia, or London in *A Study in Scarlet* to London in a *Lonely Planet* guide. It

[14] That characters' names can be associated to both diagnostic and non-diagnostic traits is similar to the characterization of concepts for modeling the history of ideas [29].

is the credit that we give to texts and the way in which we unload our stipulations that make the difference. More precisely, we individuate when a *name* is "fictional" or "non-fictional" on the basis of its conceptual and historical plausibility, i.e., by looking at its consistency with common-sense knowledge and the existence of other names with the same diagnostic traits (according to criteria (C1) and (C2)) appearing in texts for which there exists a large agreement on their historical foundation. Given the dependence on interpretations and diagnostic traits of (C1) and (C2), the individuation of the fictionality or non-fictionality of a name is interpretation and trait dependent.

A first index of non-fictionality of name n in text T is when $\texttt{TRAIT}(A,R,n,T)$ does not require the interpreters in A to renounce to common-sense knowledge (see Sect. 2). In fantasy or science-fiction worlds, common-sense often fails but note that also scientific theories may go against common-sense, e.g., the case of quantum physics. A second and more important index of non-fictionality of n in T is when one finds a name m in text U such that, according to A, there is a relation between n and m via (C1) or (C2), and U is an historically founded text according to A or even a wider community.

First, as said, these indexes are heavily dependent on A, on their interpretations of T and U, and on their selection of diagnostic traits for n and m. However, when A is a large community, a sort of intersubjective point of view on the index can be obtained. Second, by generalizing the adopted notion of text, among the texts U considered to ground the non-fictionality of n one could also include data coming from scientific experiments that usually have a high degree of intersubjectivity. Third, the agreement on the historical foundation of a text can change through time, this means that fictional or non-fictional indexes are also time dependent and subject to revision due to new discoveries.

By collecting all these indexes, a group A can at time t establish the level of fictionality of a name. For instance, intuitively, 'Sherlock Holmes' can be interpreted as a *fully fictional* name, because there is no evidence in historical texts of characters with the same traits to which scholars attribute empirical value. Names like 'Napoleon' in Tolstoy's *War and Peace* (or 'Francesca da Rimini' in Dante's *Comedy*) could be understood as *semi-fictional*, because only some of their traits can be reconducted to historical figures. The case of Boccaccio mentioned above is more subtle, since scholars do not have enough information to conclude whether some of his characters are fully fictional, semi-fictional, or historical, with the latter intended to align with a biography.

In conclusion, it should be clear that in our approach, the distinction between fictional and non-fictional names is not absolute but it rather depends on both the manner in which texts are interpreted and which diagnostic traits are associated with the names. In this sense, there could be even cases where scholars first attribute a fictional status to a certain name, whereas they may change their mind after the acquisition of empirical knowledge about it. This perspective is an important departure point with respect to the philosophical debate and in our view it remains close to literary investigations.

4. Discussion and Conclusions

The need for acquiring new means to represent and model literary texts and their interpretations is increasingly evident. This is because the debate is rich and varied from multiple sides: not only does academic literary criticism play a role, but the transmission and preservation of literary texts involve a continuous process of interpretation conducted also through online platforms such as fan blogs and services [30].

From the standpoint of literary studies, questions relative to the ontological nature of the entities texts talk about is not so central. What is pivotal instead is the way in which interpreters approach texts, e.g., what they claim, on the basis of which sources, theories for interpretation, and so on. Accordingly, an important aspect of this contribution is to lay down some key aspects for the representation of interpretations that are potentially compatible with different philosophical positions about the entities featuring in a text, be they realist or anti-realist. In addition, taking inspiration from current works in both philosophy [7] and computational literary studies [23], we have proposed to ground interpretations on inferences that interpreters make on the interpreted texts on the basis of various elements. The result of an interpreter's inference on a literary text is a (set of) sentence(s) that according to the interpreter follows from the interpreted text. We formally capture this through the mechanism of *commitment*. In line with literary studies, interpretation as commitment is primarily *subjective*, while it is still possible for multiple interpreters to share and agree on common interpretations. To develop our proposal we have relied on studies on the formal semantics of natural language and, in particular, on Discourse Representation Theory, including some of its extensions. At the same time, we have slightly adapted them, although in a preliminary fashion, to our application context to make explicit the subjectivity of interpretation. For instance, for a given text, it is up to each single interpreter to decide how to solve possible ambiguities or inconsistencies in the text, or the kind of inferences they are able to make on the basis of their subjective knowledge, or reference to other sources supporting interpretation.

We worked on a connected topic in a previous paper [5]. In that case, we required experts to share a common *observational language* in order for them to enter into a debate, that is, to share a common ontology tuned to the expression of observations. In this sense, interpretations are public commitments towards a set of propositions of an observational language partially representing the content of a text. Differently, in the present paper the only requirement for interpreters is to be competent speakers of a natural language, i.e. to be acquainted with its syntax and grammar, leaving open the possibility of having as many interpretations of the language as the scholars who participate in the debate. As said, we have assumed hereby that scholars can entertain debates, possibly agreeing on each other positions, without sharing a common ontology. Commitments are sort of assertive speech-acts that (i) do not presuppose an observational language; (ii) are moves according to the "rules of the game" of literary debates, so we do not presuppose them to be interpreted; and (iii) consider the point of view of a particular interpreter. The current proposal and the one in [5] are however compatible; e.g., the concept of *commitment* hereby explored can be used to ground the design of observational languages.

Future work to strengthen our proposal will attempt to further explore both approaches in detail, including their integration in a unifying framework in such a way to "operationalize" the modeling, comparison, and analysis of texts' interpretations.

Acknowledgments: Research leading to this work is supported by the project *MITE – Make it explicit: Documenting interpretations of literary fictions with conceptual formal models* funded by the European Union - Next Generation EU.

References

[1] Jannidis F. Character. In: The Living Handbook of Narratology. Hamburg University; 2014. Available from: http://www.lhn.uni-hamburg.de/article/character.

[2] Kroon F, Voltolini A. Fictional Entities. In: Zalta EN, Nodelman U, editors. The Stanford Encyclopedia of Philosophy. Fall 2023 ed. Metaphysics Research Lab, Stanford University; 2023. .
[3] Frow J. Character and person. Oxford University Press, USA; 2014.
[4] Galván L. Counterfactual claims about fictional characters: philosophical and literary perspectives. Journal of Literary Semantics. 2017;46(2):87-107.
[5] Sanfilippo EM, Sotgiu A, Tomazzoli G, Masolo C, Porello D, Ferrario R. Ontological Modeling of Scholarly Statements: A Case Study in Literary Criticism. In: Aussenac-Gilles N, Hahmann T, Galton A, Hedblom MM, editors. Formal Ontology in Information Systems (FOIS 2023). vol. 377 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2023. p. 349-63.
[6] Sartini B, Baroncini S, van Erp M, Tomasi F, Gangemi A. ICON: an Ontology for Comprehensive Artistic Interpretations. ACM Journal on Computing and Cultural Heritage. 2023.
[7] Paganini E. Vague fictional objects. Inquiry. 2020;63(2):158-84.
[8] Wittgenstein L. Philosophical investigations. Anscombe GEM, Hacker PMS, Schulte J, editors. London: Blackwell; 2009.
[9] Margolin U. Mathematics and narrative: A narratological perspective. In: Circles disturbed: the interplay of mathematics and narrative. Princeton University Press; 2012. p. 481-507.
[10] Arrighi C, Ferrario R. The dynamic nature of meaning. In: Magnani L, Dossena R, editors. Computing, Philosophy, and Cognition. College Publications; 2005. p. 1-18.
[11] Gius E, Jacke J. The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis. International Journal of Humanities and Arts Computing. 2017;11(2):233-54.
[12] Searle JR. Expression and meaning: Studies in the theory of speech acts. Cambridge University Press; 1979.
[13] Kamp H, Reyle U. From discourse to logic: Introduction to model-theoretic semantics of natural language, formal logic and discourse representation theory. Springer; 2013.
[14] Lascarides A, Asher N. Segmented discourse representation theory: Dynamic semantics with discourse structure. In: Computing meaning. Springer; 2007. p. 87-124.
[15] Priest G. Paraconsistent logic. In: Handbook of philosophical logic. Springer; 2002. p. 287-393.
[16] Friend S. The real foundation of fictional worlds. Australasian Journal of Philosophy. 2017;95(1):29-42.
[17] Recanati F. Mental files. Oxford University Press; 2012.
[18] De Ponte M, Korta K, Perry J. Truth without reference: The use of fictional names. Topoi. 2020;39(2):389-99.
[19] Maier E. Fictional names in psychologistic semantics. Theoretical Linguistics. 2017;43(1-2):1-45.
[20] Kamp H. Using proper names as intermediaries between labelled entity representations. Erkenntnis. 2015;80:263-312.
[21] Walton KL. Mimesis as make-believe. Harvard University Press; 1990.
[22] Masolo C, Sanfilippo EM, Ferrario R, Pierazzo E. Texts, Compositions, and Works: A Socio-Cultural Perspective on Information Entities. In: JOWO proceedings. CEUR vol.2969; 2021. .
[23] Jacke J. Is There a Context-Free Way of Understanding Texts? The Case of Structuralist Narratology. Journal of Literary Theory. 2014;8(1):118-39.
[24] Lindström S. A semantic approach to nonmonotonic reasoning: inference operations and choice. Theoria. 2022;88(3):494-528.
[25] Shoham Y. Reasoning about change: time and causation from the standpoint of artificial intelligence. Yale University; 1987.
[26] Hirsch E. Physical-object ontology, verbal disputes, and common sense. Philosophy and Phenomenological Research. 2005;70(1):67-97.
[27] Eco U. On the ontology of fictional characters: A semiotic approach. $\Sigma\eta\mu\varepsilon\iota\omega\tau\kappa\acute{\eta}$-Sign Systems Studies. 2009;37(1-2):82-98.
[28] Richardson B. Transtextual Characters. In: Eder J, Jannidis F, Schneider R, editors. Characters in Fictional Worlds: Understanding Imaginary Beings in Literature, Film, and Other Media. Berlin: De Gruyter; 2011. p. 527-41.
[29] Masolo C, Sanfilippo E, Lamé M, Pittet P. Modeling concept drift for historical research in the digital humanities. In: 1st Int. Workshop on Ontologies for Digital Humanities and their Social Analysis (WODHSA) at JOWO 2019. CEUR vol.2518; 2019. .
[30] Guillory J. Professing Criticism: Essays on the Organization of Literary Study. University of Chicago Press; 2022.