FOIS 2024

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, via alla cascata 56/c, 38123, Trento, IT*
[b] *CNR ISTC, via Gaifami 18, 95126, Catania, IT*

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 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* (ficta) in literary studies and philosophy [1,2], these including Sherlock Holmes in Doyle's stories, Emma Bovary in Flaubert's novel, Humbert Humbert in Nabokov's *Lolita*, and many others. Ficta extend to any sort of thing featuring 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 prominent role in literary studies where it is still debated if characters are pieces of writing, person-like entities or some combinations of the two [3,4].

A literary scholar may be interested in analyzing some of the traits of characters with respect to some interpretative theories, or may look at the relationship between characters across texts and authors 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 [5]. 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 interpretative theories, sources, etc.

From a computer science perspective, current efforts in Digital Humanities aim to develop systems to support scholars in their interpretation practices [4,6]. Ideally, a new

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

generation of systems 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 the same texts across cultures and epochs. With the goal of supporting the documentation and analysis of interpretations of literary texts, we present here an approach focused on interpreting characters, considering their significance in literary investigations.

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 [4]). As a consequence, we cannot simply document the way in which a text "depicts" a character, because a text must be "put in dialogue" with an interpreter to tell anything. The relationship between characters and interpretations is an important departure point with respect to the debate in analytic philosophy; it is sufficient to recall that mainstream theories on ficta do not ascribe any role to interpreters (see [2]). To the best of our knowledge, a philosophical theory emphasizing the connection between texts, characters, and interpretations was presented by Paganini [7]. In her work, Paganini 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 to consider interpreters' *dispositions* to accept the truth of interpretations. 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' dispositions are necessarily made explicit in a debate.

Second, to document interpretations, we adopt an approach based on *statements* in natural language, to which scholars publicly commit regarding 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 public statements and potential agreements among interpreters formed upon them without delving into ontological considerations. As radical as this move could seem, it is legitimate with respect to the pragmatic dimension of literary interpretation practices where experts do not 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* (PI) [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. In order 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, for example, 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; 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 [10] 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. As we will see, the approach we put forward is compatible with different perspectives on the ontological nature of ficta, realist as well as anti-realist positions.

The remaining of the paper is structured as follows. Section 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 Section 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, Section 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 [11], an updating or explication of the information contained in the text. From this perspective, the aim of interpreting a text is to clarify its content according to a group of interpreters, rather than to determine its truthfulness or factual basis.[2] We therefore assume a sort of stipulative (or pretence) attitude of interpreters [10], 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' explicit *commitments* to *public* linguistic statements. When multiple interpreters share their commitments regarding a text, we obtain a shared interpretation of the text. Following Wittgenstein [8, sect. 242], as said, this attitude is empirical: measuring (interpreting) is determined not only by sharing methods of measurement (committing to the same text), but also by constancy in results of measurement (sharing of judgments regarding the text).

[2]Empirically-based notions of "truthfulness" are relevant to establish the nature of characters, see Sect.3.

A *text* T is here understood as a sequence of *sentences* in a natural language.[3] We write $\texttt{COMMIT}(a,s,T)$ for "the interpreter a publicly commits to the truth of the declarative sentence s in the context of text T," where s and T are expressions of the same natural language and a is a competent speaker of this language. $\texttt{COMMIT}(a,s,T)$ can be understood as a sort of public *speech-act* [12] about T performed by a. The general idea behind $\texttt{COMMIT}(a,s,T)$ is that, 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 (where $T{\circ}s$ stands for the sequence of sentences obtained by adding the sentence s to the sequence T). $\texttt{COMMIT}(a,s,T)$ is therefore based on a sort of inference process, i.e., it requires a notion of truth-condition, as well as a notion of truth-preservation. However, we will show that the approach (i) is compatible with different theories of meaning, and (ii) does not presuppose a specific (shared) ontology, i.e., $\texttt{COMMIT}(a,s,T)$ and $\texttt{COMMIT}(b,s,T)$ require neither a and b to share an ontology, nor the accessibility to a of the way b is semantically grounding s (and T), and vice versa.

To better clarify our notion of commitment, 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*.[4] In these approaches, discourses are dynamically translated[5] 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. DRSs (and, indirectly, discourses) can be, in their turn, translated into first order (FO) formulas.[6]

When a discourse is (syntactically and grammatically) ambiguous and lexical or common-sense knowledge is 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. biases. 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.[7] A complication arises when texts are (superficially) logically inconsistent, e.g., when a text explicitly claims something and its negation. One can think that the shared

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

[4] The notion of discourse in DRT and SDRT is close to the notion of text previously discussed.

[5] Translation is a step-by-step process where each step can depend on the previous ones.

[6] DRT and SDRT often involve non-classical logics, particularly dynamic and non-monotonic logics. We refrain from delving into this discussion and instead assume that logic is shared among all interpreters.

[7] Possible ambiguities are represented by including into $\tau(a,T)$ the logical disjunctions of the formulas corresponding to all the DRSs associated to T accepted by a.

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 either by considering non-classical logics, or by "tolerating" them 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 inconsistency, and a form of rationality of the text's author is presupposed.[8]

As said, interpreters' commitments can be based on additional (personal or shared) knowledge. When such knowledge is inconsistent with what is reported in a (consistent) text T, our notion of interpretation (based on the above discussed pretence attitude of interpreters) presupposes that only part of such knowledge (the one consistent with T) can be used for their commitments. We indicate with $\kappa(a,T)$ the knowledge of interpreter a consistent with T and we assume that $\kappa(a,T)$ is also represented by means of FO-formulas.[9] In Sect. 2.3 we will be more specific on the knowledge of interpreters distinguishing the lexical and common-sense knowledge shared by all the interpreters from the knowledge coming from other texts explicitly considered by a to interpret T.[10]

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., what is written in T is necessary to commit to s.[11] 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 referentialist and anti-realist positions on the characters in T.

2.1. Commitments: Referentialist positions

Following standard practice in logic, the truth conditions for the formulas in $\tau(a,T)$, $\kappa(a,T)$, and $\tau(a,s)$ can be semantically grounded on set-theoretical structures. Less conventionally, we assume that an interpreter a has certain ontological commitments, i.e., a can formally interpret the language of the formulas above in terms of an intended set $\mathfrak{M}(a)$ of structures using a potentially complex interpretation function. Through this function, individual constants (predicates) are not necessarily mapped to elements (tuples of elements) of the domains of the structures in $\mathfrak{M}(a)$, but rather, an elaborated set-theoretical reduction could be necessary. $\mathfrak{M}(a)$ thus represents a third element of subjectivity in commitments.

In this framework, the previous requirement assumed for $\texttt{COMMIT}(a,s,T)$ can be 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 $\mathscr{M}\vDash\tau(a,T)\cup\kappa(a,T)$ for all $u\in\tau(a,T)\cup\kappa(a,T)$, $\mathscr{M}\vDash u$), then it also sat-

[8] In some cases, an inconsistency can be functional for the whole literary work, see for example [16].

[9] Inconsistencies with T can be solved by removing different parts of the original knoweledge of a. We ignore this issue here.

[10] For example, one may interpret a text about vampires by using knowledge from other texts about vampires.

[11] By considering $\kappa(a,T)$ among the knowledge one can use to derive $\tau(a,s)$, we embrace a pretense mediated version of the *Reality Assumption* [17] that has the known issue that everything that is in $\kappa(a,T)$ can be also the subject of the 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.

isfies all the formulas in $\tau(a,s)$ (i.e., $\mathscr{M} \vDash \tau(a,s)$). Furthermore, there exists a model $\mathscr{M} \in \mathfrak{M}(a)$ such that $\mathscr{M} \vDash \kappa(a,T)$ but $\mathscr{M} \nvDash \tau(a,s)$.

2.2. Commitments: Anti-realist positions

In a referentialist perspective, all proper names, definite descriptions, and indexicals are translated into individual constants which, in their turn, are reduced to elements of the domains of the model-theoretic structures considered by interpreters. This view has been disputed by philosophers embracing anti-realist positions with respect to ficta (see [2]). In these views, fictional names like 'Sherlock Holmes' do not refer to any entity. We will now show that we can still make sense of the previous requirement on $\texttt{COMMIT}(a,s,T)$ in line with antirealist positions by considering a psychologistic variant of the Tarskian definition of truth, where truth-conditions of fictional statements are provided in terms of interpreters' mental states.

Mental states can be represented by taking inspiration from the mechanism of *mental files* introduced by Recanati [18]. In this line of works, Korta et al. [19] distinguish different kinds of statements and provide for them different kinds of referential or non-referential truth-conditions. We provide here some more details on the approach put forward by Maier [20], because (i) it is based on an extension of DRT allowing for a more direct comparison with the previous reading of COMMIT; and (ii) it provides uniform truth-conditions for both fictional and non-fictional statements, avoiding the problem of categorizing statements under sentence-kinds subjected to different truth-conditions.

By relying on a recent work by Kamp [21], Maier [20] extends standard DRT with mental attitudes. In this approach, 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 [22], where fictional statements serve as prescriptions for imagination. In this psychologistic version of DRT, 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" [20, p.9].

Without entering into the details of the approach (see Maier [20]), the crucial aspect is that the truth-conditions for fictional and non-fictional statements are provided in terms of how a labeled-DRS (a syntactic entity) obtained from natural language statements captures (part of) an agent's mental state, which is formally represented in terms of complex structures. In this way, the truth of a statement is given in terms of the mental states of interpreters without necessarily presupposing a referent for the involved entities.[12]

2.3. Grounding commitment 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 might 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

[12] Maier assumes that the truth of certain statements can be expressed in terms of referents, enabling him to uniformly address sentences such as "Holmes lives in London" and "Holmes is a fictional character."

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 interpret literary texts *in the light of* other texts [23].

$\texttt{COMMIT}(a,s,T,U)$ stands for: "According to interpreter a, what is reported in sentence s derives 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 the previous section, $\texttt{COMMIT}(a,s,T,U)$ requires that $\tau(a,T) \cup \tau(a,U) \cup \kappa_s(a,T,U) \models \tau(a,s)$ but $\tau(a,T) \cup \kappa_s(a,T) \not\models \tau(a,s)$. Here $\kappa_s(a,T,U)$ represents the shared lexical and common-sense knowledge that is consistent 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. Analogously for $\kappa_s(a,T)$. We write $\texttt{COMMIT}(a,s,T,\emptyset)$ when no additional knowledge is required, i.e., $\tau(a,T) \cup \kappa_s(a,T) \models \tau(a,s)$ but $\kappa_s(a,T) \not\models \tau(a,s)$. For instance, assume that (1) "Holmes lives in 211 Baker Street" and (2) "Baker Street is in London" are in T. If $\kappa_s(a,T)$ 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 $\kappa_s(a,T)$ do not contain any information about the location of London, then (4) "Holmes lives in England" cannot be directly 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 that $\tau(a,T) \cup \tau(a,U) \cup \kappa_s(a,T,U)$ is consistent. This means that, to support their commitment, a considers texts that do not contradict the text T.[13] Second, to avoid to further complicate our framework, we assume a single supporting text U, but clearly a could need several texts to ground their commitment. Third, instead of grounding their commitment on what is reported in U (formalized via $\tau(a,T) \cup \tau(a,U) \cup \kappa_s(a,T,U) \models \tau(a,s)$) a could rely on some prior interpretations of the text 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. Fourth, 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. We partially analyze this aspect in Sect. 3.

2.4. From commitments to agreements and interpretations

Public 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: $\texttt{AGREE}(A,S,T,U) \coloneqq \forall a \in A(\texttt{COMMIT}(a,S,T,U))$.

[13] An interesting extension could consider possible resolutions of contradictions between T and U.

The notions of commitment and agreement can be further abstracted 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: $\texttt{COMMIT}(a,S,T) \coloneqq \exists U(\texttt{COMMIT}(a,S,T,U))$, $\texttt{AGREE}(A,S,T) \coloneqq \forall a \in A(\texttt{COMMIT}(a,S,T))$.

An interpretation of a text T is a maximal text S on which there is an agreement: $\texttt{INT}(A,S,T)$ stands for "the text S is the interpretation of the text T from the point of view of the group A." 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 question of the ontological nature of the entities we talk about becomes superfluous, i.e., the interpretation of a text is not affected by the different philosophical positions regarding the nature of these entities, be they realist or anti-realist. Analogously, Hirsch [24] observes that discussants (in philosophical debates in metaphysics) do not necessarily need to share a common ontological theory to understand each other. This is because each discussant can make sense of what others say, and vice versa, based on their private ontological theory and shared principles of conversation.

As limit cases there are personal interpretations (when A contains a single interpreter) and common interpretations (when A is the whole set of interpreters) but clearly it is possible to have different groups of interpreters agreeing on different interpretations of the same texts, e.g., $\texttt{INT}(A,R,T)$ and $\texttt{INT}(B,S,T)$ with $R \neq S$.[14]

For the sake of clarity, although we have followed Paganini's [7] idea of grounding a text's interpretation in the agreement among interpreters of the text T on the truth of a text S, our approach differs significantly from her approach. 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 certain 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 public 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 readings of COMMIT, different translations of T, U, and S, as well as different ontological commitments (especially in the case of realist positions). Furthermore, $\texttt{AGREE}(A,S,T)$ 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 a model-theoretic interpretation, since each interpreter can consider 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.

[14] 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.

3. Relations between characters' names

As we have seen in Sect. 2.3, interpreters can approach multiple texts, whether they are literary texts or other sorts of texts used to support their interpretations. In these cases, they need to understand whether the linguistic elements (names, descriptions, etc.) n and m in texts T and U, respectively, have the same meaning. Our idea is to approach this question in the light of debates around identity (see [2] for some discussions), as well as similarity or other types of relations between characters.

We consider a simplified scenario here, 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 on additional agreements about certain sentences assumed to characterize the meaning of such names (this point will be clarified later).[15] Third, while a given name may have different meanings in different texts, we assume that a name maintains the same meaning within a single text.

3.1. Diagnostic traits

Among the traits that scholars consider to identify and analyze characters, they might focus on a subset of traits considered as particularly relevant. To report an example, discussing about the character of Emma Bovary, Eco [25] claims that her character in Gustave Flaubert's novel and Woody Allen's film script for the *The Kugelmass Episode* is the same character, 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 [26]. As a less extreme example, in the case of *series* of texts like the stories of Doyle, intuitively Holmes remains the same character even though across the books of the series he has quite different traits and just maintains the relevant ones. Let us then assume that characters are associated with *diagnostic traits* (to adopt Eco's terminology) used to identify them across texts. We will explore how this approach can be represented and exploited in our framework.

In our model, we take into account characters by considering their names and what is said about them that, given our interpretive stance, have interpretation-dependent meanings. Interpreters may commit to multiple sentences involving a given name n in T, while considering only some of them as salient for n. In other words, interpreters select the relevant sentences for n in T, corresponding to the diagnostic traits, from among those in their interpretations of T. Considering our notion of interpretation, it is not entirely clear to us whether this selection process is purely interpretative, meaning whether the relevance of a sentence for n can be derived from T and lexical/common-sense knowledge (unless explicitly stated in T).

[15] 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.

For these reasons, we prefer to introduce a new kind of public commitment: $\texttt{sTRAIT}(a,R,n,T)$ stands for "the interpreter a publicly declares that the sentences in R are all relevant for the name n as appearing in the text T",[16] where $\texttt{sTRAIT}(a,R,n,T) \rightarrow \texttt{COMMIT}(a,R,T)$, i.e., the relevant sentences for n in T are included in a's interpretation of T. Following what done for the interpretation, $\texttt{TRAIT}(A,R,n,T)$ collects in R all the relevant sentences for n in T on which the group A of interpreters agree (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 only the point of view of the group A of interpreters. Said that, (C1), where $V_{(n_1 \rightarrow n_2)}$ indicates the text obtained by syntactically substituting in V the name n_1 with the name n_2, assures that (a) A selected the diagnostic traits for both n in T and m in U; and (b) 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 group 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 $a \in A$, $\kappa_s(a,T) \models \tau(a,R) \leftrightarrow \tau(a,S_{(m \rightarrow n)})$ and $\kappa_s(a,U) \models \tau(a,S) \leftrightarrow \tau(a,R_{(n \rightarrow m)})$.

C2 According to the group 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) $\texttt{AGREE}(A,R_{(n \rightarrow m)},U,\emptyset)$.

(C1) and (C2) (as well as their weaker versions) establish links between names as appearing in given 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. To follow philosophical creationism on fictional entities [27] one may however easily restrict the previous criteria to texts with the same author. Further refinements can be introduced when additional information about the time at which texts have been produced or about the explicit reference of an author to previous texts is available.[17]

(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.[18] 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 of 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 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

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

[17] This additional information can be controversial and founded on sources not always recognized by experts.

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

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, along with philosophical possibilism, one may assume that when inconsistent traits are identified for a fictional name, different characters have to be distinguished. These may stand in counterfactual modal relations, rather than identity, while having only consistent traits.

From this latter perspective, one might assume that (C1) (or (C2)) does not imply identity of meaning but weaker relations, e.g., what one may call *borrowing*. For instance, one can say that Emma Bovary in Allen's work is borrowed from Emma Bovary in Flaubert's work, because the diagnostic traits are preserved and there is an explicit intention of Allen to refer to Flaubert's work. In this case, we have two non-interchangeable names, i.e., intuitively, Flaubert's Emma Bovary and Allen's Emma Bovary are different characters. The recognition of authors' *intentions* to borrow characters from other texts can be problematic in a literary perspective, especially without evidence in written sources on which scholars may rely. Borrowing can be in its turn weakened into *derivation* by considering the weaker versions of (C1) (and (C2)), i.e., derivation requires only an (partial) overlap between the traits associated to the names that may not even capture diagnostic traits.[19]

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 somehow nuanced from a literary standpoint. For some characters like Holmes or Emma Bovary, scholars commonly 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 a young woman called Francesca da Polenta happened during Dante's life. However, there remain only few historical traces about the latter, 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 to this landscape, there are plenty of figures for which scholars do not know whether they are fictional or historical, like in the case of Boccaccio's text *De mulieribus claris* (14th century).

In our approach the boundary between "reality" and "fiction" can be traced based on the relationship between texts (see [26] 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 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.

[19] That characters' names can be associated to both diagnostic and non-diagnostic (marginal) traits is similar to the characterization of concepts for modeling the history of ideas [28].

An 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 can go against common-sense, e.g., quantum physics.

A second, and probably more important, index of non-fictionality of n in T is when it is possible to find a name m in text U such that, according to A, n in T and m in U are linked via (C1) or (C2) and U is a text on which at least A, but in general a wider community, agrees on its historical foundation. 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, even though a modification of the adopted notion of text is required, 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 documented in sources with empirical value. 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 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

Nowadays, the need for acquiring new means to represent and model a crucial aspect of human creativity – narration and its interpretations – is increasingly evident. This is because the debate is rich and varied: not only does academic literary criticism play a role, but also the transmission and preservation of literary works involve a continuous process of interpretation, conducted by scholars, critics, and, more recently, online platforms such as blogs and services like Goodreads [29].

We have worked on a connected topic in a previous paper [4]. The present work may be seen as a foundation of it, which does presuppose neither a commitment on the nature of fictional characters and, more generally, on an ontology shared by all interpreters, nor the necessity to express interpretations via a shared *observational language*.

As said throughout the paper, from the empirical but also normative standpoint of literary studies the question of the ontological nature of the entities we talk about becomes in a certain sense superfluous. For us, this is not to avoid these types of questions. Instead, it means that we aim at a representation of the interpretative activity that allows different philosophical positions regarding ficta, be they realist or anti-realist, to subsist without this affecting the interpretation of texts.

In the perspective put forward in [4], we required experts to share a common vocabulary for an observational language in order for them to enter into a debate. 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. Hence, in this current work we deepen the study of concepts such as commitment and agreement that can be eventually used to found the design of observational languages (we leave this to future work). In [4] the interpretations of texts are expressed via *assertions*, i.e., public statements about the fact that a piece of information is "contained" in a specific text and what is asserted is a proposition of an observational language. Commitments are then sort of assertions that however (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. Notice that this last aspect can be considered in the approach in [4] by "reifying" the commitment itself, i.e., by requiring it to be included in a text as well. Since assertions can be nested, i.e. they can range over assertions that in their turn range over other assertions, commitments can be seen as assertions of assertions, that is, assertions stated in the text that reifies the commitment of assertions about the original text (that however are limited to propositions of the observational languages, i.e., they are not simply natural language sentences as in commitments). A further common line of both works is the use of the notion of *source* to document the origin of a piece of knowledge or information. In [4] the source indicates the text in which an observation may be found, while in this work with "source" we refer to an additional text that, together with the text to be interpreted (and possibly with background knowledge), allows the interpreter to infer new information, in the form of a sentence.

We would like to conclude this paper by pointing out that the study of the textual sources of information is becoming unavoidable, especially after the advent of Large Language Models and generative AI, whose reliability in the production of trustworthy output is debatable or – to tell the least – difficult to evaluate [30]. One of the purposes of the project at the basis of this work is exactly that of documenting the *interpretative game*, by checking the commitments of every interpretation and the pattern of agreements involving it, in search of possible signs of problems (for instance, an interpretation involved in few agreements might be considered problematic). Our hope is that this approach could mitigate the issues of trustworthiness that can emerge with LLMs and generative AI in general.

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] 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.

[5] Swirski P. Literature, analytically speaking: Explorations in the theory of interpretation, analytic aesthetics, and evolution. University of Texas Press; 2010.

[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. 2019.

[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: Doxiadis A, Mazur B, editors. Circles disturbed: the interplay of mathematics and narrative. Princeton University Press; 2012. p. 481-507.

[10] Galván L. Counterfactual claims about fictional characters: philosophical and literary perspectives. Journal of Literary Semantics. 2017;46(2):87-107.

[11] 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.

[12] Searle J. What is a speech act? In: philosophy in America. Routledge; 2014. p. 221-39.

[13] Kamp H, Reyle U. From discourse to logic: Introduction to model-theoretic semantics of natural language, formal logic and discourse representation theory. vol. 42. 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] Phillips JF. Truth and inference in fiction. Philosophical Studies. 1999;94(3):273-93.

[17] Friend S. The real foundation of fictional worlds. Australasian Journal of Philosophy. 2017;95(1):29-42.

[18] Recanati F. Mental files. Oxford University Press; 2012.

[19] De Ponte M, Korta K, Perry J. Truth without reference: The use of fictional names. Topoi. 2020;39(2):389-99.

[20] Maier E. Fictional names in psychologistic semantics. Theoretical Linguistics. 2017;43(1-2):1-45.

[21] Kamp H. Using proper names as intermediaries between labelled entity representations. Erkenntnis. 2015;80:263-312.

[22] Walton KL. Mimesis as make-believe. Harvard University Press; 1990.

[23] 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. .

[24] Hirsch E. Physical-object ontology, verbal disputes, and common sense. Philosophy and Phenomenological Research. 2005;70(1):67-97.

[25] Eco U, et al. 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.

[26] 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.

[27] Thomasson AL. Fiction and metaphysics. Cambridge University Press; 1999.

[28] Betti A, Van den Berg H. Modelling the history of ideas. British Journal for the History of Philosophy. 2014;22(4):812-35.

[29] Guillory J. Professing Criticism: Essays on the Organization of Literary Study. University of Chicago Press; 2022.

[30] Chaturvedi A, Bhar S, Saha S, Garain U, Asher N. Analyzing Semantic Faithfulness of Language Models via Input Intervention on Question Answering. Computational Linguistics. 2024 02:1-37.