Journal for the History of Analytical Philosophy
Volume 15, Number 4 (2026)

Putnam's Semantic Externalism Revisited

Hiroto Takagi
takagi.hiroto.1125@gmail.com
Graduate School of Letters,
Kyoto University

ABSTRACT

From "Is Semantics Possible?" (1970) to his seminal work "The Meaning of 'Meaning'" (1975), Hilary Putnam developed his semantic externalism about the meaning and reference of natural kind terms. His metasemantic position, especially his idea of 'indexicality', is typically interpreted as a form of physical externalism. Such a view is committed to both natural kind realism (as a basis for reference determination) and causal-historical chains (as a basis for reference preservation). I contend that this interpretation requires reconsideration. Some scholars have presented textual evidence that Putnam did not presuppose natural kind realism when he developed his metasemantic position during the relevant period. Nonetheless, the question of how Putnam's semantic externalism should be reformulated remains open. In this paper, I reconstruct his metasemantic position as what I term convergent externalism. On this account, the diachronic identity of reference is both guided by norms inherent to our linguistic practices and retroactively secured. While Putnam revised his general philosophical views throughout his career, my thesis demonstrates that the metasemantic and epistemological foundations of his thought were more unified than is often recognized.

1. Introduction

Between "Is Semantics Possible?" (Putnam 1970) and his much-discussed "The Meaning of 'Meaning'" (Putnam 1975a), Hilary Putnam developed a metasemantic theory that has come to be known as *semantic externalism*.

This view denies semantic internalism, the position that the meaning or reference of a natural kind term[1] supervenes on a speaker's intrinsic

[1] Although semantic externalism itself is not confined to natural kind terms and Putnam claims that this view is applied to artificial kind terms like "pencil" (Putnam 1975a, 242–45),

properties. Putnam famously expressed this idea by asserting that "'meanings' just ain't in the *head*" (Putnam 1975a, 227, emphasis in original). On semantic externalism, the meaning of the natural kind terms a speaker uses is, at least partially, determined by her physical and/or social environment.

Putnam's semantic externalism, especially his idea of "indexicality", is commonly interpreted as a form of *physical externalism*,[2] which is consistent with the view Saul Kripke develops in *Naming and Necessity* (Kripke 1980).[3] On this interpretation, Putnam adopts these two claims: (1) natural kinds as a mind-independent division of the world (i.e., natural kind realism) determine the reference of natural kind terms and (2) the reference is transmitted from person to person via a causal-historical chain. Despite being widely accepted, this interpretation is worth reconsidering. Indeed, several scholars have argued that Putnam did not presuppose natural kind realism when he developed his metasemantic position between 1970 and 1975 (Ebbs 1992; Hacking 2007, 2015; Wikforss 2013).

However, the question of how Putnam's semantic externalism should be reformulated remains open. In this paper, I aim to reinterpret Putnam's semantic externalism as what I term *convergent externalism*. This reinterpretation also suggests that there is more continuity in Putnam's intellectual development than is often recognized.[4] Putnam is well-known for having changed his mind about his general philosophical views. His intellectual career is often characterized as a transition from *realism* (metaphysical realism, Early Putnam) to *antirealism* (internal realism, Middle Putnam) and then back to *realism* (natural realism, Late Putnam) (see Baghramian 2008; Macarthur 2020). On this reading,

this paper focuses on natural kind terms. As for the recent discussion about Putnam's claim about artificial kind terms and semantic externalism, see Bianchi (2022).

[2]The term "physical externalism" is used in Wikforss (2013) and Hickey (2009). Other studies refer to Putnam's semantic externalism as a "causal theory of natural kind terms" (Devitt and Sterelny 1976) or "natural kind externalism" (Kallestrup 2013). These terms all refer to physical externalism.

[3]For an explanation of the difference between Putnam's and Kripke's metasemantic positions, see Hacking (2007) and Wikforss (2013).

[4]Few have attempted to find unity within Putnam's thought. Nevertheless, Ebbs (1992) and Hickey (2009) argue that Putnam's philosophical project should be understood as an attempt to explicate the norms of the socio-linguistic practices in which we participate. I mostly agree with this characterization.

Putnam's semantic externalism is tied to his metaphysical realism phase, at which time he held that reality is divided into objects in a determinate way, independent of human activity (see Baghramian 2008, 19). I intend to reconstruct Putnam's semantic externalism in a way that does not entail a commitment to real divisions of natural kinds. In so doing, I shall both (a) present a new interpretation of Putnam's semantic externalism and (b) shed new light on his intellectual career.

This paper is structured as follows. In Section 2, I provide a brief introduction to Putnam's semantic externalism. In Sections 3 and 4, I outline and criticize the physical externalism interpretation. In Section 5, I reconstruct how Putnam attempted to explain the diachronic identity of the reference of natural kind terms by reformulating what he called "Principle of Benefit of Doubt". In Section 6, I elaborate on this principle by examining how Putnam justified his position. Finally, in Section 7, I discuss some implications of my interpretation, which can help us better understand Putnam's intellectual journey.

2. An Outline of Putnam's Semantic Externalism

In this section, I introduce the concept of indexicality, which plays a central role in Putnam's semantic externalism. After "Is Semantics Possible?", Putnam consistently argued against what he referred to as 'traditional theories about meaning'.[5] In "The Meaning of 'Meaning'", he presents both his own metasemantic position and a *reductio ad absurdum* against his opponents. I shall discuss this topic in this section.

In "The Meaning of 'Meaning'", Putnam characterizes traditional theories about meaning as relying on the following two assumptions (Putnam 1975a, 219, 222):

> (TM1) The psychological state of an individual speaker determines the meaning of a term.

[5] Along with "Is Semantics Possible?" and "The Meaning of 'Meaning'", Putnam also discussed semantic externalism in "Explanation and Reference" (Putnam 1973a), "Meaning and Reference" (Putnam 1973b), "Comment on Wilfrid Sellars" (Putnam 1974a) and "Language and Reality" (Putnam 1974b). According to Takaaki Matsui, Putnam's idea of the division of linguistic labor and the Twin Earth thought experiment were already present in "Comment on Wilfrid Sellars" (see Matsui 2021, sec. 3).

> (TM2) The meaning of a term (that is, its intension or Fregean 'sense') determines its reference in the sense that sameness of intension entails sameness of reference.

Putnam argues that (TM1) is based on an assumption he calls 'methodological solipsism'. This assumption states that a subject's psychological state is solely determined by her intrinsic properties and not by the external environment. Therefore, according to (TM1), the meaning of a natural kind term is determined by what the speaker mentally associates with that term. For instance, 'lemon' can be defined as an "object with properties such as a sour taste, thick yellow peel, elliptical shape, and yellow pulp". According to (TM2), the meaning of a natural kind term—defined as per (TM1)—functions as its intension, which serves as the necessary and sufficient condition for its reference (Putnam 1975a, 220–22).

Putnam then presents a *reductio ad absurdum* against traditional theories about meaning. According to the argument's structure, (TM1) and (TM2) jointly entail the following claim:

> (TM3) The psychological state of an individual speaker determines the reference of a natural kind term she utters.

Putnam argues that (TM3) is false, and, therefore, either (TM1) or (TM2) must also be false. Traditional theories about meaning maintain that if two speakers are in the same mental state, then a natural kind term uttered by each speaker will have the same reference. However, Putnam contends that this is not always the case. There are cases in which the descriptions a speaker mentally associates with a particular natural kind term fail to determine the reference. To illustrate this, Putnam presents two thought experiments: (1) the Elm and Beech case and (2) the Twin Earth thought experiment. These examples are designed to demonstrate that (TM3) is false.

Let me begin with the Elm and Beech case. Putnam states that he cannot distinguish between an Elm tree and a Beech tree. This is because he associates the same concepts with these natural kind terms; he does not know the differences between them. At best, he might link them to the concept "tree" or "deciduous tree". Nevertheless, he insists that when he says "elm", he refers to elm, and when he says "beech", he refers to beech (Putnam 1975a, 227). In other words, the references of natural kind terms such as "elm" and "beech" differ in his language despite

being associated with the same descriptions. According to Putnam, the "division of linguistic labor" in a linguistic community helps to address this issue when descriptions are insufficient to determine the reference of a natural kind term. The division of linguistic labor refers to a social practice in which laypeople defer to experts within their linguistic community—experts who can distinguish the references of relevant natural kind terms (Putnam 1975a, 228).[6]

Putnam's other thought experiment—the famous Twin Earth thought experiment—proceeds as follows: Imagine a planet called Twin Earth located somewhere in the galaxy. Twin Earth is a perfect physical and historical replica of Earth, with one exception: the substance referred to as "water" on Twin Earth is not the same substance as the water on our Earth. Instead, it is an alien substance whose molecular structure is abbreviated as "XYZ". XYZ behaves identically to water at normal temperature and pressure. It is odorless, transparent, and tasteless (among other properties). Putnam predicts that, if we Earthlings today discovered that twin water is XYZ, we would revise our initial supposition that 'water' on Twin Earth means water as understood on Earth. Putnam further asks us to imagine that there is a pair of molecular duplicates on Earth and Twin Earth in 1750: Oscar on Earth and T-Oscar on Twin Earth. In the 18th century, chemistry was not yet fully developed; even experts did not clearly understand the chemical structures of water and twin water. This means that the relevant sociolinguistic context could not determine the reference of the term "water" in 1750. Therefore, Oscar and T-Oscar naturally associated the same descriptions with 'water' while the references were different. According to Putnam, this thought experiment demonstrates that the references of natural kind terms are not "a function of the psychological state of the speaker" (Putnam 1975a, 224).

It is, however, unclear how we can claim that the reference of "water" on Earth in 1750 is the same as it is today, which means that the reference of "water" on Earth was distinct from that of "water" on Twin Earth, even in 1750. In this scenario, we cannot rely on the division of linguistic labor. This is because (1) even the experts could not discern water from twin water in 1750 and (2) the division of linguistic labor is a synchronic

[6]For further discussion on the division of linguistic labour and its relevance for recent discussions on linguistic deference, see Moldovan (2016).

relation between experts and laypeople, not a diachronic one. As Lance Hickey points out, the Twin Earth thought experiment is a case where "the division of linguistic labor would be of no help to individuate the reference of 'water' on Earth from Twin Earth" (Hickey 2009, 22). Therefore, to explain the diachronic identity of the references of natural kind terms, Putnam is required to appeal to a theoretical device other than the division of linguistic labor.

Putnam's answer is that the reference of the term "water" on Earth remains the same over time because natural kind terms have "an unnoticed indexical component" or *indexicality* (Putnam 1975a, 233–34). Indexicals—such as "I", "now", "this", or "here"—change their reference depending on context. Suppose that Oscar and T-Oscar simultaneously say, "I have a headache". The reference of "I" will differ depending on who is speaking. Even if Oscar and T-Oscar are perfect duplicates in the same psychological state, the "I" Oscar utters refers only to Oscar and the "I" T-Oscar utters refers only to T-Oscar. Putnam argues that natural kind terms behave like indexicals. The references of natural kind terms are sets of substances that bear *sameness relations* with local samples of each kind. As such, the references change from one physical environment to another (Putnam 1975a, 225). Furthermore, the sameness relation of a natural kind is determined by its "important physical properties", which are typically intrinsic properties such as chemical structure (Putnam 1975a, 238–39). This is why Putnam argues that water and twin water cannot be the same, even in 1750. According to Putnam, the reference of "water" remains constant over time (as does that of "water" on Twin Earth).

3. An Outline of the Physical Externalism Interpretation

In this section, I shall introduce the standard interpretation of Putnam's semantic externalism. On this interpretation, Putnam's idea of indexicality is properly understood as a form of physical externalism, also known as the causal theory of reference.[7] I shall refer to this as "the PE interpretation". I will argue against this interpretation in Section 4.

[7]For the sake of simplifying the discussion, "physical externalism" and the "causal theory of reference" will not include the "causal-descriptive theory", according to which

As described in (Devitt and Sterelny 1976), physical externalism consists of two components: (1) the theory of reference fixing and (2) the theory of reference borrowing. The former explains how a natural kind term obtains its reference, while the latter describes how the reference is transferred (or preserved) from speaker to speaker. An "[o]stensively introduced sample and natural kind realism" and the "causal-historical chain" correspond to reference fixing and reference borrowing, respectively (Devitt and Sterelny 1976, 70–71).

Regarding reference fixing, physical externalism incorporates the notion of a *dubbing ceremony*.[8] At this ceremony, a natural kind term is introduced, and its reference is determined. The original dubber selects a specific substance and associates it with a specific name stating something like, "From now on, we call this 'water' ". The reference of a natural kind term is determined by a sameness relation (referred to as "being of the same kind" in Devitt and Sterelny (1976, 70)). This relation obtains between the ostensively introduced sample and other members of the kind (Devitt and Sterelny 1976, 70).

Regarding sameness relations, the PE interpretation maintains that Putnam presupposes natural kind realism. It is beyond the scope of this paper to compare various definitions of this position. I shall, therefore, adopt the general formulation found in Tahko (2015, 796):

> Natural Kind Realism:
> There are entities—the natural kinds—which reflect natural divisions in mind-independent reality.

The PE interpretation holds that Putnam assumes that the membership of each natural kind reflects a pre-existent or built-in division, meaning that it does not depend on the categorization systems that we introduce, and that the real division of natural kinds determines the reference of natural kind terms. As a result, "users of a natural kind term need not . . . know the necessary and sufficient conditions for membership" (Devitt and Sterelny 1976, 70–71).

Some philosophers of language (e.g., Devitt and Sterelny 1976; Kallestrup 2013; Wolf 2002) and some philosophers of science (e.g.,

descriptions play a role in reference fixing. For more details on this position, see Devitt and Sterelny (1976, 91–92) and Crane (2021, sec. 3).

[8] This is equivalent to what Putnam referred to as an "introducing event" (Putnam 1973b, 200).

LaPorte 2004; Khalidi 2023) attribute natural kind realism to Putnam. Even Putnam scholars commonly assume that he presupposed natural kind realism (or scientific realism) when he wrote his semantic externalist papers (see Baghramian 2008; Hickey 2009; Macarthur 2020). Hickey, for example, interprets the central claim of the Twin Earth thought experiment to be that the reference of natural kind terms "is determined not by the sociolinguistic community but by *the nature of the world itself*" (Hickey 2009, 22, emphasis added). He adds that "[t]his has sometimes been called "physical" or even "metaphysical externalism" to distinguish it from the social externalism derived from a mere appeal to the division of linguistic labor" (Hickey 2009, 22).

On the other hand, the theory of reference borrowing explains how individual speakers can refer to a natural kind when they lack sufficient knowledge to discern it. According to the PE interpretation, after the reference of a natural kind term is fixed at the *introducing event*, the reference is then transmitted to other speakers via causal links, such as communication and education. These links allow speakers to refer directly to natural kinds even if they do not causally interact with them and/or lack the conceptual resources to specify them (Devitt and Sterelny 1976, 71).

In summary, the PE interpretation maintains that Putnam's response to the question "why should we accept that the term 'water' has the same reference in 1750 and 1950 (on both Earths)?" (Putnam 1975a, 224) is as follows: A dubber associated a sample of water with the term "water" in the past. The term gained its reference through the world's built-in categorization. The reference is then preserved over time through a causal-historical chain.

4. The Case Against the PE Interpretation

Since the publication of "The Meaning of 'Meaning'", philosophers have examined Putnam's semantic externalism from various perspectives. John McDowell, for instance, has attempted to elucidate the consequences of Putnam's metasemantic stance on the character of mental content (McDowell 1992). Neil Williams argues that Putnam's view of natural kinds should be construed as a form of neo-essentialism, according to

which the essences of natural kinds are not limited to their intrinsic properties (Williams 2011).[9]

However, as stated, few scholars have questioned the idea that Putnam presupposed natural kind realism when he wrote about semantic externalism. Contrary to the conventional understanding, I aim to demonstrate in this section that Putnam's semantic externalism did not postulate natural kind realism during the relevant period. (Note that even if Putnam's semantic externalism did not presuppose natural kind realism, this fact does not entail that his entire philosophical position during the relevant period abandoned natural kind realism. This paper is only about his position on natural kind *terms*, not on natural kinds themselves).[10]

On the PE interpretation (and on physical externalism), the sameness relation of a natural kind is independent of the human categorizations we cast onto the world. Ian Hacking has argued against this interpretation. He notes that Putnam's question "what is the relation same_{L} [i.e., the relation of being the same liquid]?" (Putnam 1975a, 239) does not imply a commitment to natural kind realism (Hacking 2007, 9). Putnam states as follows:

> x bears the relation same_{L} to y just in case (1) x and y are both liquids, and (2) x and y agree in important physical properties. . . . What I focus on now is the notion of *importance*. Importance is an interest-relative notion. Normally the 'Important' properties of a liquid or solid, etc., are the ones that are *structurally* important: the ones that specify what the liquid or solid, etc., is ultimately made out of – elementary particles, or hydrogen and oxygen, or earth, air, fire, water or whatever – and how they are arranged or combined to produce the superficial characteristics. (Putnam 1975a, 238–39, emphasis in original)

Thus, Putnam believes that the sameness relation is determined by "important physical properties". Furthermore, the matter of which properties count as important properties is *relative to our interests*. Internal structures (e.g., water's molecular composition, H_2O) establish the sameness relation as long as they serve our interests. For instance, one of the key aims of contemporary scientific inquiry is to elucidate the causal factors underlying the observable behavior of a given natural

[9] Sanford Goldberg and Andrew Pessin's *The Twin Earth Chronicles* (Goldberg and Pessin eds. 2016) is a collection of significant papers on Putnam's semantic externalism.

[10] Thank you to an anonymous reviewer for clarifying this point.

phenomenon. It is not merely about identifying correlations between these behaviors. Thus, the reference of a natural kind term is dependent on subjects' interest.

Putnam discusses a case in which a sameness relation is not identified with a specific structural property due to the interests of the classifying subjects. In China, the term "jade" refers to two distinct minerals: jadeite and nephrite. These minerals have different molecular structures. Jadeite consists of sodium and aluminum, while nephrite comprises calcium, magnesium, and iron. Modern chemists, therefore, regard them as separate substances. However, both materials can be shaped through abrasion rather than carving (resulting in exquisite forms) (Hacking 2015, 349). The Chinese value this specific characteristic and use the same name to refer to both jadeite and nephrite.[11] As this case shows, according to Putnam, the members of a natural kind do not "necessarily *have* a common hidden structure" and the sameness relation can depend on the possession of superficial characteristics when the identification serves the interest of the subjects (Putnam 1975a, 240–41, emphasis in original).

Even in "Is Semantics Possible?", which Putnam describes as "my first explicitly 'semantic externalist paper'" (Putnam 2016, 202), he did not adopt natural kind realism. There, he argued against traditional theories of meaning, contending that natural kinds had "explanatory importance" because observable characteristics "are 'held together' or even explained by deep-lying mechanisms" or "essential nature" (Putnam 1970, 139–40). Due to Putnam's emphasis on essence, natural kind realism—particularly traditional natural kind essentialism, which holds that mind-independent natural kinds are defined by essential and intrinsic properties (Tahko 2015, 796; Williams 2011, 151)—is often attributed to Putnam's semantic externalism.

However, his emphasis is placed more on theories of natural kinds than on natural kinds themselves.

> [T]he knowledge of the properties that a thing has (in any natural and non 'ad hoc' sense of property) is not enough to determine . . . whether or not it is a lemon (or an acid, or whatever). . . . Meaning does not determine reference, in the sense that given the meaning and a list of all the 'properties' of a thing (in any particular sense of 'property') one can

[11] See Hacking (2015) and LaPorte (2004) for a more detailed discussion and the history of the term "jade".

> simply *read off* whether the thing is a lemon (or acid, or whatever). Even given the meaning, whether something is a lemon or not, is, or at least sometimes is, or at least may sometimes be, a matter of what is the best conceptual scheme,[12] the best theory, the best scheme of 'natural kinds'. (Putnam 1970, 142)

In the quoted paragraph, Putnam argues that descriptions associated with a natural kind term cannot, by themselves, fix the reference, even if the descriptions include details about microstructures. But what does fix the reference of a natural kind term? Putnam's answer is that it is the best theory of natural kinds. In sum, what Putnam argues against traditional theories is not that natural kinds themselves fix the references of natural kind terms. Rather, he argues that the inclusion criteria for a natural kind term is constantly modified through the course of scientific development and, therefore, concepts associated with a natural kind term at a certain period cannot fix the reference of the natural kind term. This is why he states, "*today we would* say it [the essential nature] was chromosome structure, in the case of lemons, and being a proton-donor, in the case of acids" (Putnam 1970, 141, emphasis added). His emphasis on internal structure merely describes the behavior of today's scientists, as seen similarly in "The Meaning of 'Meaning'".

However, there are still two questions that should be addressed. First, if Putnam's semantic externalism does not presuppose natural kind realism, as shown above, how does he explain the diachronic identity of the reference of natural kind terms? In other words, his idea of indexicality needs to be reformulated. Second, what does Putnam mean by "essential nature" or "theory-independent entities," which evoke the notion of a built-in categorization of reality? The first question will be addressed in the next section, and the second question in Section 6.

5. Indexicality and Deference

Several scholars have pointed out that Putnam's semantic externalism did not presuppose natural kind realism (Ebbs 1992; Hacking 2007, 2015;

[12] Hickey states that Putnam first used the term "conceptual scheme" in *Meaning and Moral Sciences* (Hickey 2009, 60). This passage shows that the term already appeared in the 1970 paper "Is Semantics Possible?".

Wikforss 2013). However, none have reformulated the idea of indexicality, which is central to Putnam's defense of the diachronic identity of the reference of natural kind terms. In other words, it remains unclear how Putnam answers the question "why should we accept that the term 'water' has the same reference in 1750 and in 1950 (on both Earths)?" (Putnam 1975a, 224). In this section, I aim to (re)interpret the notion of indexicality.

Let us begin by reviewing the Twin Earth thought experiment. As stated in Section 2 and as Wikforss (2013, 249–50) rightly notes, Putnam's supposition in this thought experiment is that people in 1950 would revise the initial judgement that "water" has the same meaning on Earth and Twin Earth when they discover that the substance called 'water' on Twin Earth is XYZ (Putnam 1975a, 223). He further states that "the reference of the term 'water' was just as much H_2O on Earth in 1750 as in 1950" (Putnam 1975a, 224). Moreover, Oscar and T-Oscar "understood the term 'water' differently in 1750 although they were in the same psychological state" (Putnam 1975a, 224). As mentioned in the previous section, Putnam also considered the reference fixing of natural kind terms to be interest-relative and theory-dependent. These facts suggest that Putnam believed that the reference of "water" posited in 1950 was *retroactively* attributed to "water" uttered in 1750 and, therefore, "water" had the same reference in both time periods.

Putnam is explicitly committed to retroactive reference attribution in his paper "Language and Reality" (Putnam 1974b). There, he discusses the diachronic identity of reference of theoretical terms such as "quark" and "electron". He argues that the reference is secured by a sort of *semantic deference* to today's scientific experts. Semantic deference is a phenomenon where a word's reference is determined by the usage of others, particularly suitable experts. According to Putnam, scientists often introduce theoretical terms using descriptions intended to single out the relevant theoretical entities. These descriptions are sometimes misdescriptions, but this does not mean that they fail to refer to the relevant entities. Scientists in the past can be understood as referring to entities posited in today's scientific theories by invoking what Putnam calls "The Principle of Benefit of Doubt" (PBD). He writes:

> The Principle of Benefit of Doubt is simply the principle that we should give the dubber, or the relevant expert, if the person at the other end

> of the chain of transmissions or cooperation is not the original dubber, the benefit of doubt in such cases by assuming that he would accept reasonable modifications of his description. (Putnam 1974b, 275)

PBD, also known as the "presumption of innocence", is a legal convention in criminal trials. It requires the court to deliver a not guilty verdict unless the prosecution has established its case beyond a reasonable doubt. Putnam extends this principle to the context of the reference of theoretical terms. He argues that experts in the past must be judged to have successfully referred to some theoretical entity unless there is reasonable doubt that their reference failed. In other words, a verdict of reference failure should be established beyond a reasonable doubt. The reasonable doubt in question relates to the assumption that experts in the past would accept and defer to scientific theories in the future (because the latter are "reasonable modifications" of the former). In this case, the reference is retrospectively attributed to past experts by assuming that they were enough rational to acknowledge their fallibility.

In "The Meaning of 'Meaning'", Putnam similarly explains the diachronic identity of reference of natural kind terms in terms of the semantic deference of earlier-day scientists. He uses the example of gold and an identical-looking substance, X. While X can be easily distinguished from gold using contemporary chemistry, it could not be distinguished using techniques in Ancient Greece. Putnam claims that "'gold' has not changed its reference ... in two thousand years. ... [T]he reference of $\chi\rho\upsilon\sigma\acute{o}\varsigma$ in Archimedes' dialect of Greek is the same as the reference of *gold* in my dialect of English" (Putnam 1975a, 235). Putnam provides three reasons to support this claim: First, when Archimedes claimed that a substance was gold, he supposed that (a) the substance shared certain observable characteristics with gold and (b) it had the same 'hidden structure' as (any other local samples of) gold. Second, we assume that Archimedes would defer to contemporary scientists. Although he could not differentiate between gold and X, if experiments were performed in which X behaved differently from gold, he would conclude that X is not gold. Third, Archimedes would agree with today's scientists that X is not gold if they informed him about the two substances' differing molecular structures. Thus, Putnam attempts to explain the diachronic identity of the reference of natural kind terms

by appealing to a supposed semantic deference to current scientific experts.

Two points are worth noting here. First, we should avoid interpreting Putnam's talk of an "essential nature" or "mind-independent entities" as a commitment to natural kind realism. Such an interpretation assumes that semantic deference serves as a proxy for deferring to the built-in categorization of the world itself.[13] Second, Putnam's view does not require that people in the past explicitly defer to a term's future usage. Rather, it requires us only to suppose that they *would* accept future expert modifications of their usage.

In summary, on Putnam's semantic externalism, the reference of natural kind terms is fixed by the most rationally acceptable theory available rather than by the structure of reality.[14] Putnam's theory also suggests that reference preservation is retroactive, meaning that it is not a causal-historical relationship. Instead, preservation of reference is achieved by assuming that past experts were sufficiently rational to acknowledge their fallibility and defer to future experts. Note also that both reference fixing and reference preserving are normative relationships in Putnam's semantic externalism.

6. Putnamian Temporal Externalism

The view I attributed to Putnam in the last section represents a form of temporal externalism. Following Henry Jackman, this view holds that "we typically understand ourselves as taking part in a shared, temporally extended, and *ongoing* practice" (Jackman 1999, 158, emphasis in original). This suggests that reference of linguistic expressions can be partly determined by usage in the future.

In this section, I shall explicate Putnamian temporal externalism by addressing two questions: The first question concerns what Putnam

[13] See Section 1 of De Brabanter and Leclercq (2023) for an explanation of the relationship between various types of semantic externalism and semantic deference.

[14] Note that, although the built-in structure of the world does not fix the references of natural kind terms, the external world contributes to fixing them within Putnam's semantic externalism. As he puts it, "[t]raditional semantic theory leaves out only two contributions to the determination of extension—the contribution of society and *the contribution of the real world*!" (Putnam 1975a, 245, emphasis added). In other words, new information and findings can lead to modifications in the theory and, consequently, in the references of natural kind terms.

means by terms such as "essential nature", "hidden structure", and "theory-independent entities" given that these terms remind us of an unconceptualized reality to which his metasemantic position does not presuppose. The second question is how Putnam tries to justify his assumption about past speakers' deference to future usage (i.e., PBD). I will address these questions in turn.

Let us look at a passage where Putnam characterizes "theory-independent entities" at some length. He writes:

> It is beyond question that scientists use terms as if the associated criteria were not *necessary and sufficient conditions*, but rather *approximately* correct characterizations of some world of theory-independent entities, and that they talk as if later theories in a mature science were, in general, *better* descriptions of the same entities that earlier theories referred to. (Putnam 1975a, 237, emphasis in original)

In this paragraph, Putnam discusses how scientists view earlier theories. According to him, scientists "talk as if later theories in a mature science were, in general, better descriptions of the same entities that earlier theories referred to" and the existence of this linguistic practice is "beyond question". The key phrase here is "as if". Putnam's claim can be restated as follows: When new evidence leads to modifications in the extension of a natural kind posited in a scientific theory, the earlier theory is *reinterpreted as an imprecise description of entities posited in the later theory*.[15, 16] Also, our best current theories are potentially open to reconstruction in the sense that today's scientific experts will be interpreted according to PBD in the future. Such a possibility of modification is expressed as "theory-independent entities" or "essential nature". This answers our first question.

Now, let us turn to our second question. This question asks how Putnam attempted to justify the assumption that past speakers would have deferred to future usage, i.e., the legitimacy of PBD. Putnam seems to provide two justifications for PBD in the following paragraph:

[15] This is why Putnam calls a sameness relation of a natural kind "a defeasible necessary and sufficient condition" (Putnam 1975a, 225).

[16] Putnam's view does not always require the reconstruction of past theories as predecessors of later theories (because he incorporates the phrase "in general" in the quoted passage). We do not think that 'phlogiston' refers to valance electrons, for example (Putnam 1988, 14). Rather, the term 'phlogiston' simply does not have a referent.

> Like all methodological principles it [PBD] is partly a descriptive principle; I assume that we all wish the benefit of doubt to be accorded to us when we are the dubbers and the experts–thus the principle describes intentions which actually exist and are for the most part honored in the linguistic community–and it is a *normative* principle; we should honor it, for otherwise stable reference to theoretical entities would almost surely be impossible (Putnam 1974b, 275, emphasis in original)

In this paragraph, Putnam highlights PBD's dual nature: it is both descriptive and normative. The principle is descriptive because it describes our desire to be interpreted as sufficiently rational to defer to an expert successor. In other words, we wish to be seen as fallible but simultaneously *on the path towards convergence to our successor theories and contributing to scientific development*.

This marks the difference between Putnam's approach and how contemporary research often attempts to justify temporal externalism. Take Jussi Haukioja's work as an example of recent research in this area. In "Semantic Burden-Shifting and Temporal Externalism", he introduces the notions of "metainternalism" and "metaexternalism". Metainternalism and metaexternalism provide answers to the question of what makes a specific semantic internalist or externalist view correct. According to metainternalists, the correct theory is determined by factors internal to the speaker at the time of utterance. In contrast, metaexternalists reject this view (Haukioja 2020, 922).

Haukioja advocates for metainternalism as a justification for temporal externalism. He appeals to what he calls "burden-shifting dispositions", which are "second-order dispositions to re-evaluate and retract one's applications of a word in response to new information" (Haukioja 2020, 923). He argues that temporal externalism is true of (at least) natural kind terms because "speakers are disposed to accept and go along with a range of different interpretations of the term they have introduced, depending on how future speakers decide to use the term" (Haukioja 2020, 926).[17] In contrast, Putnam's argument for metainternalism–as a justification for temporal externalism—is based on the idea that we want to be interpreted as fallible while also converging toward (successful) successor theories and contributing to scientific development. In this way,

[17] Jackman, who originated temporal externalism, also seems committed to metainternalism when he states as follows: "When we make an utterance, we often commit ourselves to future refinements in communal usage . . ." (Jackman 1999, 160).

Putnam's and Haukioja's views are subtly different. I shall, therefore, refer to Putnam's semantic externalism as *convergent externalism*.

Putnam's attempt to justify his convergent externalism also includes a metaexternalist component, although he does not claim it to be *true* but rather considers it *obligatory* from a metaexternalist perspective. According to Putnam, PBD is not merely descriptive but also a normative principle that we should follow (Putnam 1974b, 275). The reason he provides is that, without PBD, we could not compare theories and, consequently, could not ensure "a growth of objective knowledge" (Putnam 1974b, 281; see also Putnam 1975a, 235–38). Putnam's argument lacks elaboration. Nonetheless, one could argue that if the reference of natural kind terms were determined solely by their associated descriptions, then different theories would assign different referents to the same term. In such cases, comparison between them would be impossible because they would pertain to different subject matters—a phenomenon which is often called "incommensurability". This is understandable, because Putnam, as mentioned earlier, does not assume an inherent categorization of the world. In this sense, PBD functions as a regulative assumption—an assumption that we must adopt in order to engage in rational inquiry (Howat 2013, 453; Misak 2013, xi). Putnam's justification for PBD is best paraphrased as follows: adopting PBD is obligatory because, without it, we could not ensure the possibility of scientific progress in the first place and would therefore have no reason to engage in scientific inquiry. This constitutes the metaexternalist component of convergent externalism.

7. Further Discussion

I have presented a new interpretation of Putnam's semantic externalism. Before concluding this paper, I shall briefly discuss the implications this interpretation has for understanding his intellectual career.

As mentioned, Putnam is often thought to have radically changed his general philosophical views throughout his career. The most famous "turn" was his supposed move to internal realism in his 1976 John Locke Lecture (Putnam 1978). Following convention, I shall refer to Putnam before 1976 as 'Early Putnam' and Putnam after 1976 as "Middle Putnam".

Middle Putnam[18] is often categorized as a *neopragmatist* (alongside Richard Rorty, W. V. O. Quine, and Robert Brandom).[19] Neopragmatists reject the idea that we can step outside our conceptual framework and determine whether it corresponds to the world as it is. Quine explains as follows: ". . . we cannot detach ourselves from [our conceptual scheme] and compare it objectively with an unconceptualized reality. Hence it is meaningless, I suggest, to inquire into the absolute correctness of a conceptual scheme as a mirror of reality" (Quine 1950, 632).

Middle Putnam shared this view, arguing that "[t]here is no God's Eye point of view that we can know or usually imagine" and that "there are only the various points of view of actual persons reflecting various interests and purposes that their descriptions and theories subserve" (Putnam 1981, 50). Based on this philosophical approach, he characterized one of his philosophical projects as "to find a picture that enables us to make sense of the phenomena [of intentionality] *from within our world and our practice*, rather than to seek a God's-Eye View" (Putnam 1988, 109, emphasis added). Indeed, he maintained that one of classical pragmatism's key insights is "the thesis that . . . practice is primary in philosophy" (Putnam 1994, 152) [20] and he himself emphasized what he called "the agent point of view" (Putnam 1987, 70; 1992, 351).

As mentioned, Early Putnam–with his semantic externalism–is often linked to a commitment to the natural categorization of the world as independent of human activities. He is, consequently, seldom linked to pragmatism. In her monograph on the history of American pragmatism, Cheryl Misak notes that "[i]n his early work, he had written important papers in philosophy of science, metaphysics, and philosophy of mathematics" but "[t]here was not much in that material to signal what would by 1980 be a major shift towards pragmatism" (Misak 2013, 238).

Nonetheless, a remark in his autobiography suggests that Putnam was already a pragmatist when he wrote "The Meaning of 'Meaning'":

[18] According to Putnam, ""Meaning Holism" contains some improvements I regard as important, and I wish it better known" (Putnam 2015, 78). However, the purpose of this paper is to reconstruct his semantic externalism in 70's so I do not address "Meaning Holism" (Putnam 1986) here.

[19] The term "neopragmatism" is not strictly defined. In this context, it is used to distinguish Putnam (and Rorty, Quine, and Brandom) from classical pragmatists such as C. S. Peirce, William James, and John Dewey.

[20] The other insights of classical pragmatism that Putnam raises are fallibilism, antiskepticism, and the denial of the fact/value dichotomy (Putnam 1994, 152).

> In short, I saw myself as describing and, to a certain extent, reconstructing, the practices–e.g., the division of linguistic labor–that are presupposed by our talk of meaning intertheoretically at all. ... Indeed, some years later, in a paper read to Montreal World Congress in philosophy, Kripke expressed dissatisfaction with the "The Meaning of 'Meaning'" precisely on the ground that the notion of the 'essence' of a natural kind I employ is not independent of scientific practice. (Putnam 1992, 349)

There are two points in this paragraph worth noting. First, Putnam does not object to Kripke's charge. This suggests that Putnam's semantic externalism in "The Meaning of 'Meaning'" does not presuppose natural kinds as independent of scientific practice, a point substantiated in the discussion in Section 4. Second, Putnam states that he aimed to reconstruct the linguistic *practice* necessary for discussing the meaning of natural kind terms across different theories. If the PBD is included in "the practices ... that are presupposed by our talk of meaning intertheoretically at all", as mentioned in the quoted paragraph, the emphasis on "practice" and "the agent point of view," which are normally attributed to Middle Putnam, had already appeared in Early Putnam's works. Therefore, the convergent externalism interpretation suggests greater coherence in the metasemantic and epistemological foundations of Putnam's philosophical stance than is commonly assumed, indicating that he was already a neopragmatist when he developed semantic externalism.[21,22]

8. Conclusion

To conclude, in the first half of the paper (Sections 2 and 3), I outlined the notion of indexicality, which is central to Putnam's semantic externalism. I also introduced the PE interpretation, according to which indexicality includes (a) natural kind realism as a basis for reference fixing and (b) a causal-historical chain as a basis for reference preservation. In the

[21] Putnam's much-discussed no-miracles argument in "What is Mathematical Truth?" (Putnam 1975b) was published in the same year as "The Meaning of 'Meaning'". This argument is often considered to be an argument for scientific realism. If Early Putnam was indeed a neopragmatist, then the no-miracles argument might need to be reconsidered.

[22] Pierre-Yves Rochefort has discussed Putnam's supposed transition from internal realism to common sense realism in the 1990s. Rochefort points out salient similarities between these two positions (Rochefort 2021).

second half, I presented a new understanding of Putnam's metasemantic position. Following several prior studies, I demonstrated that Putnam's semantic externalism was not committed to natural kind realism. Instead, he argued that natural kinds are theory-dependent entities (Section 4). I also interpreted Putnam's semantic externalism as what I called *convergent externalism* (in accordance with PBD). This means that the reference of natural kind terms is determined by future theories and that reference is then retroactively preserved (Section 5). I then discussed Putnam's justification for convergent externalism. He maintained that PBD reflects our actual intentions and represents the only way to avoid incommensurability (Section 6). These discussions suggest a new interpretation. My argument points to more continuity–more unity–in the metasemantic and epistemological foundations of Putnam's thought than scholars usually assume (Section 7).

Acknowledgements

I am grateful to Takaaki Matsui, Michael Campbell, and Satoshi Kodama for helpful comments on earlier drafts. Drafts of this article were presented at Kyoto University and The Japan Forum for Young Philosophers in 2023. Thank you to the audiences for helpful feedback. My work has been financially supported by JST SPRING, grant number JPMJSP 2110.

References

Baghramian, Maria. 2008. "'From Realism Back to Realism': Putnam's Long Journey." *Philosophical Topics* 36: 17–35.

Bianchi, Andrea. 2022. "Kind Terms and Semantic Uniformity." *Philosophia* 50: 7–17.

Crane, Judith. 2021. "Two Approaches to Natural Kinds." *Synthese* 199: 12177–98.

De Brabanter, Philippe, and Benjamin Leclercq. 2023. "From Semantic Deference to Semantic Externalism to Metasemantic Disagreement." *Topoi* 42: 1039–50.

Deutsch, Mathias. 2023. "Is There a 'Qua Problem' for a Purely Causal Account of Reference Grounding?" *Erkenntnis* 88: 1807–24.

Devitt, Michael, and Kim Sterelny. 1976. *Language and Reality: An Introduction to the Philosophy of Language*. Oxford: Basil Blackwell.

Ebbs, Gary. 1992. "Realism and Rational Inquiry." *Philosophical Topics* 20: 1–33.

Goldberg, Sanford, and Andrew Pessin, eds. 1996. *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "the Meaning of 'Meaning'"*. Armonk, NY: M. E. Sharpe.

Hacking, Ian. 2007. "Putnam's Theory of Natural Kinds and Their Names Is Not the Same as Kripke's." *Principia* 11: 1–24.

———. 2015. "Natural Kinds, Hidden Structures, and Pragmatic Instincts." In *The Philosophy of Hilary Putnam*, edited by R. E. Auxier, D. R. Anderson, and L. E. Hahn, 337–64. Chicago: Open Court.

Haukioja, Jussi. 2020. "Semantic Burden-Shifting and Temporal Externalism." *Inquiry* 63: 919–29.

Hickey, Lance. 2009. *Hilary Putnam*. New York: Continuum.

Howat, Andrew. 2013. "Regulative Assumptions, Hinge Propositions and the Peircean Conception of Truth." *Erkenntnis* 78: 451–68. Original text had typo 'Erkenttnis', corrected to standard spelling.

Jackman, Henry. 1999. "We Live Forwards but Understand Backwards: Linguistic Practice and Future Behavior." *Pacific Philosophical Quarterly* 80: 157–77.

Kallestrup, Jesper. 2013. *Semantic Externalism*. London: Routledge.

Khalidi, Muhammad Ali. 2023. *Natural Kinds*. Cambridge: Cambridge University Press.

Kripke, Saul. 1980. *Naming and Necessity*. Oxford: Blackwell.

LaPorte, Joseph. 2004. *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Macarthur, David. 2020. "Exploding the Realism–Antirealism Debate: Putnam Contra Putnam." *Monist* 103: 370–80.

Matsui, Takeshi. 2021. "Inferentialism and Semantic Externalism: A Neglected Debate Between Sellers and Putnam." *British Journal for the History of Philosophy* 20: 126–45.

McDowell, John. 1992. "Putnam on Mind and Meaning." *Philosophical Topics* 20: 35–48.

Misak, Cheryl. 2013. *The American Pragmatism*. Oxford: Oxford University Press.

Moldovan, Andrei. 2016. "Deference and Stereotypes." *European Journal of Analytic Philosophy* 12: 55–71.

Putnam, Hilary. 1970. "Is Semantics Possible?" In *Mind, Language, and Reality: Philosophical Papers, Volume 2*, 139–52. Cambridge University Press.

———. 1973a. "Explanation and Meaning." In *Mind, Language, and Reality: Philosophical Papers, Volume 2*, 196–214. Cambridge University Press.

———. 1973b. "Meaning and Reference." *The Journal of Philosophy* 70: 699–711.

———. 1974a. "Comment on Wilfrid Sellers." *Synthese* 27: 445–55.

———. 1974b. "Language and Reality." In *Mind, Language, and Reality: Philosophical Papers, Volume 2*, 272–90. Cambridge University Press.

———. 1975a. "The Meaning of 'Meaning'." In *Mind, Language, and Reality: Philosophical Papers, Volume 2*, 215–71. Cambridge University Press.
———. 1975b. "What Is Mathematical Truth?" In *Mathematics, Matter, and Methods: Philosophical Papers, Volume 1*, 60–78. Cambridge: Cambridge University Press.
———. 1978. *Meaning and Moral Sciences*. London: Routledge and Kegan Paul.
———. 1980. *Mind, Language, and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press.
———. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.
———. 1986. "Meaning Holism." In *Realism with a Human Face*, edited by James Conant, 278–302. Cambridge, MA: Harvard University Press.
———. 1987. *The Many Faces of Realism*. Chicago: Open Court.
———. 1988. *Representation and Reality*. Cambridge, MA: MIT Press.
———. 1992. "Replies." *Philosophical Topics* 20: 347–408.
———. 1994. "Pragmatism and Moral Objectivity." In *Words and Life*, edited by James Conant, 151–81. Cambridge, MA: Harvard University Press.
———. 2015. "Intellectual Autobiography." In *The Philosophy of Hilary Putnam*, edited by R. E. Auxier, D. R. Anderson, and L. E. Hahn, 3–110. Chicago: Open Court.
———. 2016. "The Development of Externalist Semantics." In *Naturalism, Realism and Normativity*, edited by Mario De Caro, 199–213. Cambridge, MA: Harvard University Press.
Quine, W. V. O. 1950. "Identity, Ostension and Hypostasis." *The Journal of Philosophy* 47: 621–33.
Rochefort, Pierre. 2021. "Did Putnam Really Abandon Internal Realism in the 1990s?" *European Journal of Pragmatism and American Philosophy* 13: 1–16.
Tahko, Tuomas E. 2015. "Natural Kind Essentialism Revisited." *Mind* 124: 795–822.
Wikforss, Åsa. 2013. "Bachelor, Energy, Cats and Water: Putnam on Kinds and Kind Terms." *Theoria* 79: 242–61.
Williams, Neil. 2011. "Putnam's Traditional Neo-Essentialism." *The Philosophical Quarterly* 79: 151–70.
Wolf, Michael. 2002. "Kripke, Putnam and the Introduction of Natural Kind Terms." *Acta Analytica* 17: 151–70.

Journal for the History of Analytical Philosophy

VOLUME 15, NUMBER 4 (2026)

ISSN 2159-0303

jhaponline.org