applied sciences

MDPI

Article

Implementing Cognitive Semantics of Autoepistemic Membership Statements: The Case of Categories with Prototypes

Radosław Piotr Katarzyniak [1], Grzegorz Popek [1] and Marcin Żurawski [2,*]

[1] Faculty of Information and Communication Technology, Wrocław University of Science and Technology , Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland; radoslaw.katarzyniak@pwr.edu.pl (R.P.K.); grzegorz.popek@pwr.edu.pl (G.P.)
[2] Polytechnic Faculty, University of Kalisz, Nowy Świat 4, 62-800 Kalisz, Poland
* Correspondence: m.zurawski@uniwersytetkaliski.edu.pl

Abstract: This article presents a model of an architecture of an artificial cognitive agent that performs the function of generating autoepistemic membership statements used to communicate beliefs about the belonging of an observed external object to a category with a prototype. The meaning of statements is described within the model by means of cognitive semantics. The presented proposal builds upon a pre-existing architecture and a semantic model designed for a simpler case of categories without a prototype. The main conclusion is that it is possible to develop an interactive cognitive agent capable of learning about categories with prototypes and producing autoepistemic membership statements fulfilling requirements of Rosch's standard version of prototype semantics and satisfying pragmatic and logical rules for generating equivalents of these statements in natural languages. Detailed results include the following: an original proposal for an agent's architecture, a model of an agent's strategy of learning categories with a prototype, a scheme for determining the computational complexity of particular implementations of the learning strategy, definitions of cognitive semantics for particular cases of autoepistemic membership statements, and an analytical verification of properties of the proposed cognitive semantics. Finally, this article discusses the directions of further development and potential variants of the proposed architecture.

Keywords: agent architecture; natural language generation; cognitive semantics; autoepistemic modality; embodied ontology; categorization; prototype theory

Citation: Katarzyniak, R.P.; Popek, G.; Żurawski, M. Implementing Cognitive Semantics of Autoepistemic Membership Statements: The Case of Categories with Prototypes. *Appl. Sci.* **2024**, *14*, 1609. https://doi.org/10.3390/app14041609

Academic Editor: Yi Guan

Received: 16 October 2023
Revised: 7 February 2024
Accepted: 14 February 2024
Published: 17 February 2024

1. Introduction

1.1. Cognitive Semantics

This article presents models and methods for managing the process of generating autoepistemic statements about the belonging of observed objects to conceptual categories with a prototype. It continues, and at the same time, builds upon previous results from a broad research project aiming at a design of a set of technical implementations of cognitive semantics for a variety of classes of autoepistemic modal statements (see Section 2 below). The project assumes that in its final form an application of cognitive semantics is going to functionally enable software agents to meaningfully participate in semantic communication using natural or semi-natural language.

Research on cognitive semantics of language statements originates from a field of cognitive linguistics. A fundamental goal is to formulate a detailed description of a relationship between linguistic (external, realized using a graphic or verbal form) representations of internal states exhibited (experienced) by a language-processing entity (in this case, an artificial cognitive agent, referred to as the agent) and cognitive structures reflecting an internal non-linguistic representation of said states within the entity. A formulation of a model of such a relationship for a particular statement ϕ (of natural or semi-natural language) is equivalent to a definition of cognitive semantics of this statement. A model

of cognitive semantics of the statement needs to specify elements of the agent's internal state which further serve as requirements for a generation of the statement in its linguistic (external) form. Cognitive semantics defined as such is understood as a description of the statement's meaning which is intrinsic to the agent generating the statement.

An idea of cognitive semantics can be naturally understood using the semiotic triangle which is a theoretical tool originally derived from the field of semiotics [1,2]. It helps to distinguish cognitive semantics from other types of semantics (e.g., classic set theory-based semantics). The semiotic triangle (presented in Figure 1) reflects mutual relationships present in languages of semantic communication between the following:

- symbols—Simple or compound elements of the semantic language, external to the agent's mind,
- thoughts or references—Elements of the agent's mind corresponding to these symbols;
- referents—Broadly understood objects of the external world.

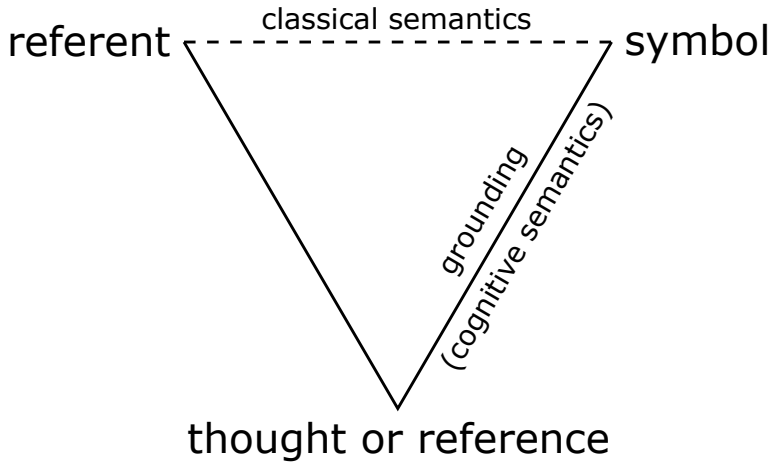

Figure 1. Semiotic triangle.

Particular implementations of the semiotic triangle are always formed in relation to a particular cognitive agent and with respect to a particular language of semantic communication. In basic implementations of the semiotic triangle, referents are usually physical objects located in the real world and symbols, in a graphical or verbal form, are interpreted as tools used to point at particular aspects of these objects. Such implementations can be found in research on a controlled societal development of the language, in particular, the naming game (numerous works of Steels, Vogt, and others, e.g., [3–6]), with theoretical ideas dating back to 1950s and Wittgenstein's works on language games.

Cognitive semantics within a semiotic triangle model binds a symbol with a corresponding reference and establishes an intrinsic (in this case, internally experienced) meaning of the symbol. It not only specifies elements of the subjective experience of the knowledge entity that corresponds to the external referent, but further describes a way in which the entity refers to parts of this experience during further processing.

For a particular sign, cognitive semantics describes a structure and contents of the internal reference related to the external sign. Signs with an assigned cognitive semantics become fully functional elements (statements) of a language of semantic communication and can be used by the agent to describe its convictions about perceived objects of the real world. In consequence, from the agent's point of view, signs assigned with cognitive semantics become so-called symbols.

An idea of cognitive semantics outlined above aligns with a sense assigned within a broad and heterogeneous field of cognitive linguistics (compare [7]). Although originally cognitive semantics has been analyzed for natural languages, it is worth pointing out that over the years some elements of this approach to the modeling of meaning have been successfully applied in research on the semantics of communication between animals [8].

Real-life implementations of the semiotic triangle for a natural language in artificial cognitive agents try to mimic ways in which human minds store acquired empirical experience and ways in which they (often deliberately and creatively) refer to this experience in order to communicate parts of its contents to others. The reference, even for a case of the simplest natural language statements, becomes a very complex system consisting of multiple mental elements. In consequence, a design (modeling) of cognitive semantics for particular types of natural language statements is a task of substantial theoretical complexity as it involves references to multiple elements of an artificial mind.

1.2. The Considered Case of Autoepistemic Membership Statements

The cognitive semantics proposed in this paper is designed and analyzed for so-called autoepistemic membership statements which have two fundamental pragmatic goals assigned in natural languages. Namely, the agent generating a particular autoepistemic membership statement fulfills a compound intention:

- to focus an attention of the recipient on its mental experience correlated with a relation of membership of a particular real or abstract object to a particular category of objects,
- to inform the recipient about a level of agent's own confidence about the compliance of a communicated belief (about the object membership to the category) with the real state of affairs.

Obviously, such a statement must meet the grammar constraints of the semantic communication language used by both the sender and recipient, be composed of at least three dedicated syntactic elements indicating an object, a category, and an experienced confidence level. It is naturally implied that both the sender and a potential recipient are required to know the pragmatics, semantics, and syntax of the language.

Due to the specificity of the research results described in this article, some additional introductory comments are required on the following two issues: the nature of the considered conceptual categories and a permissible number of confidence levels mentally distinguishable by agents involved in acts of semantic communication.

In regard to the nature of the categories, the research reported in this article concerns the so-called categories with a prototype, which differentiates this research from our previous studies into cognitive semantics of autonomous membership statements covering the case of categories described by the so-called classical theory of categorization. The latter theory has been widely recognized for many centuries, based on the ideas developed by Aristotle, and often referred to as the *necessary and sufficient condition model* [9] or *criterial-attribute model* [10]. Its basic assumptions are [11] as follows:

- shared properties—there are necessary and sufficient conditions (features) for belonging to a category; each element of a category has all of these characteristics, and there is no element outside the category that has all of these characteristics,
- clear boundaries—it is possible to unequivocally determine whether an item belongs to a category or not; this is in line with the classical set theory,
- uniformity—all elements of the category are equal, there are no more and less important elements; likewise, no distinction is made between the importance of the conditions of belonging to a category,
- inflexibility—the boundaries of the categories do not change.

Results from research conducted in the second half of the twentieth century, in virtually all cognitive science subdisciplines, showed that in many cases the above assumptions were too strict, and in consequence, were not met. For example, Geeraerts [12] pointed out that for a category *bird*, it is impossible to find necessary and sufficient conditions that are valid for all types of birds. Indeed, if we assume that a bird is an animal that is oviparous and has a beak, it turns out that there are species (e.g., a platypus) or even orders of animals (e.g., turtles) that posses these features but are not birds. On the other hand, the features that seem distinctive are missing from some members of the category: ostriches and penguins cannot fly, kiwis do not have wings, and penguins do not have the typical feathers.

Research conducted by Rosch [13,14] showed that in many cases people did not treat all elements of a particular category equally and some elements were considered more representative for the category than others. In particular, the systematic and extensive experiments conducted by Rosch proved that for different objects belonging to a category, the categorization time, the time after which the element is categorized in the learning process, and the prototypicality rating obtained from participants of the experiments could differ significantly. Rosch presented a proposal of systematic and integrated interpretation of the above effects, and summarized them as the so-called standard version of the prototype semantics.

Following [9], the basic assumptions related to categories with a prototype are usually summarized as follows:

1. The category has an internal prototype structure.
2. The degree of representativeness of a given item corresponds to the degree of its membership to a category.
3. The elements of a given category do not have properties common to all elements; they are connected by family resemblances.
4. The boundaries of categories or concepts are fuzzy.
5. The belonging to a given category is based on the degree of similarity to the prototype.
6. The belonging to a category is not determined in an analytical manner, but rather in a holistic manner.

In our research, we assume that the above assumptions on categories with prototypes must be adequately mapped into the originally formulated definition of cognitive semantics and effectively used in implemented software agents. It is worth pointing out that a process of category formation (learning) is interesting from a broader perspective as prototype-based approaches in general, and categories with prototypes in particular, are used in a plethora of knowledge processing tasks. Among others, Wang et al. [15] use a prototype-based approach for intent perception; Yongjie et al. [16] advocate a prototype-based approach to alleviate problems with outliers during unsupervised domain adaptation; Zhou et al. [17] introduce a class prototype discovery method and show further promising results in unsupervised domain adaptation.

In regard to confidence levels, it is assumed in this article (following previous research) that agents are capable of mentally distinguishing three levels of confidence about the compliance of the communicated belief (about the object's membership to a category) with the real state of affairs. Namely, an agent can be in one of the following states:

- a state of its "mind" in which it fully believes in an object belonging to the category,
- a state in which such belief is not full but still substantially intense,
- a state in which such belief is not absent at all but is significantly reduced in intensity.

These three states could be seen as certainty, strong belief, and weak belief. An introduction of three distinct levels of confidence is reflected in the pragmatics and semantics of multiple classes of natural languages. Among others, it is observable in English and Polish, both equipped with dedicated autoepistemic operators (or other means) for labeling rather three, and not more or less, levels of confidence.

Obviously, our final model of cognitive semantics, including a sub-model of reference (or thought) forming the semiotic triangle as part of it, will have to take into account the most important and necessary aspects of mental experience related to the processing of prototypical effects and to the capability of differentiating between assumed levels of autoepistemic confidence. It will also include a model of empirical knowledge base supporting the actual cognitive differentiation of levels of confidence in particular real situations.

Concluding the preliminary remarks on the considered case of autoepistemic membership statements, it can be said that in the subsequent parts of this article, the above-mentioned pragmatical, theoretical, and commonsense-related interpretations will be applied to statements for which the general structure can be summarized by the following symbol:

$$\Sigma(x \in c), \tag{1}$$

and the correlated symbol:

$$\Sigma(x \notin c), \tag{2}$$

where $\Sigma, x, c, \in, \notin$ are assigned the following interpretation: Σ is a label of an internally experienced level of confidence, x is a pointer to an internally located representation of object, c is a name of an internally represented category with prototype, $\in$ and $\notin$ are labels denoting belonging and non-belonging of x to c. In such a perspective, $\Sigma, x, c, \in$, and $\notin$ are to be understood as "physically" (e.g., graphically or verbally) realized pointers to particular elements (aspects) of an internally located complex and usually, as will be shown later, multidimensional reference.

1.3. The Main Research Target and Outline of This Article

This article introduces elements of technical implementations of the cognitive semantics and corresponding computational methods. From a broader perspective, it should be seen as an applied research work aiming at an application of models and methods of cognitive linguistics designed for modeling the meaning of symbols for semantic languages. These models are now being applied in a context of technical systems with linguistic capabilities, e.g., interactive AI systems.

In a more detailed and specific approach, the main research target of this article is to propose and at least partially verify an original version of a model of a software agent with a competence to process autoepistemic modal statements with the above-characterized cognitive semantics and pragmatics. This goal has led to the formulation of the following set of crucial functionalities required by the cognitive agent:

- implementation of a mental space of objects with a corresponding ability to tell the objects apart using an effective evaluation of the designed function of cognitive similarity (or distance) of objects,
- a formal model of a category specifying a prototype of the category and an "area" of applicability of the category,
- an internal learning process—a mechanism for an autonomous acquisition of a category model with a prototype.

The organization of the rest of this article is as follows. The next section presents a brief review of related research works which influenced the final shape of the described research project. Due to the nature of the project, it includes chosen papers from the authors' research group on other classes of autoepistemic modal statements, as well as papers setting a broader theoretical background for the original approach to modeling the meaning of the languages of semantic communication.

The third section is devoted to a more detailed discussion of the syntax of the considered class of modal statements and the commonsense (intuitive) meaning attributed to these statements. In other words, the section is devoted to an enumerative review of the basic syntactic elements that comprise the modal statements considered in this article, and to a discussion of the relationship of these elements with their equivalents in natural languages.

The fourth section contains a detailed discussion of the software agent model. The software agent ultimately acts as an entity that uses the original semantics of autoepistemic modal statements proposed in this article and manages the collection of categories with a prototype available to the agent. The agent model includes, in particular, the basic empirical database used in the process of acquiring and updating models of categories with a prototype.

The fifth section is devoted to a detailed presentation of an original strategy for learning categories with a prototype. This section considers a version of the strategy based on a universe of objects that is paired with a measure of distance as a tool for the technical realization of the agent's mental ability to compare objects. An extended computational example using an assumed distance function is provided for convenience in this section.

The sixth and seventh sections are devoted to the presentation of original cognitive semantics and to the discussion of its properties. Suitable illustrative computational examples are included.

The last two sections contain the summary of this article and point at some of our further research goals.

2. Related Work

This article proposes management tools for cognitive semantics of autoepistemic membership statements for categories with prototypes. It constitutes a distinct step of the long research project aiming at a definition, design, analysis, and application of cognitive semantics for a broad variety of classes of autoepistemic modal statements. The goal is to apply the proposed cognitive semantics in software agents in order to enable their meaningful participation in a semantic communication using natural or semi-natural language. Each class of statements already analyzed in the project reflected a particular class of autoepistemic modal statements from a particular natural language. A range of research has been similar for each class of such statements and has covered the following specific research tasks:

- a definition of a model of an agent serving as a managing subject of a respective cognitive semantics,
- a definition of an original model of a cognitive semantics within a context of an assumed agent,
- a definition of a set of postulates representing the pragmatics of using statements from a considered class in natural language,
- a formal theoretical analysis of a feasibility of the model's parametrisation guaranteeing a fulfillment of assumed postulates (which, for an assumed case, leads to a proof of a respective thesis that, for a given class of autoepistemic statements, it is possible to implement a consistent, commonsense, and interpretable set of communicative behaviors of the assumed agent).

In all of the analyzed cases, the agent communicated three distinct levels of autoepistemic confidence of belief.

The most similar case has been presented in [18] which provided the cognitive semantics of autoepistemic membership statements with a general syntactic structure given as $\Sigma(x \in c)$ or $\Sigma(x \notin c)$ for what was formally identical to the case investigated in this article. The crucial difference lies in a fact that article [18] assumed categories without prototypes. Still, it showed that for that relatively similar case it is possible to parameterize a model of an agent in a way guaranteeing a commonsense interpretation and consistency of the agent's language behaviors.

Analogous results [19–21] have been presented for autoepistemic modal statements $\Sigma(x \in p \wedge x \in q)$ which are modal extensions of conjunctions of membership statements, again for the case of categories p and q without prototypes. The actual scope of these works also included the remaining three cases of conjunction, i.e. $\Sigma(x \in p \wedge x \notin q)$, $\Sigma(x \notin p \wedge x \in q)$, and $\Sigma(x \notin p \wedge x \notin q)$. Similar remarks apply to the other research works mentioned in the current paragraph. A case of disjunctions $\Sigma(x \in p \vee x \in q)$ and $\Sigma(x \in p \veebar x \in q)$ covered in [22,23] has brought to attention an interesting pragmatic-theoretic aspect of a difference between real-language usage of inclusive and exclusive disjunction in comparison to a typical understanding of these connectives in classic formal logic. Some introductory studies into the cognitive semantics of autoepistemic implications $\Sigma(x \in p \Rightarrow x \in q)$ were proposed in [24,25]. Works [26,27] on modal equivalences $\Sigma(x \in p \Leftrightarrow x \in q)$ for categories without prototypes covered an additional aspect of the obligatory verification of the scope of processing of the collected empirical experience by the agent, and in particular the experience determining whether objects belong to a category. The feasibility of the approach has been shown both through analytical proofs and using simulations.

One of the implications of the above works is that when it comes to natural languages, even for a relatively simple set of autoepistemic statements dealing with an inclusion of an object to a particular category (in previous works, a category without a prototype), a corresponding reference exhibits a multi-level composite nature. As expected, the complexity of a reference grows after categories without a prototype are replaced with categories with a prototype.

From a broader perspective, management models for cognitive semantics of autoepistemic membership statements introduced in this article are strongly correlated with important threads of both theoretical and applied research related to a representation and processing of the meaning of natural language statements within artificial cognitive systems. The following references are worth noting.

Firstly, the approach to the modeling of cognitive semantics presented in this article follows an assumption of bipolarity of natural language symbols which, as a feature, is underlined in multiple fundamental models describing symbols of communication languages. Bipolarity is clearly reflected within the semiotic triangle presented before. The presence of two poles of a symbol (in some theories referred to as a sign), which is constituted in the form of a reference interwoven into mental structures of the subject of knowledge on the one hand, and externally realized using a correlated material form (e.g., graphical or verbal) on the other hand, can be found, among others, in the basic work by de Saussure [28] (see the distinction between *signifiant* and *signifié*), in cognitive grammar [29,30] (see the concept of semantic unit consisting of *phonological unit* and *symbolic unit*), as *internal patterns* and *external patterns* in neurobiological interpretation of meaning [31,32], as well as in the more classic dual coding approach to mental representations proposed by Paivio [33] (see *mental images and image memory* versus *verbal representations and verbal memory*).

All of the above approaches present a natural bipolarity of the symbols as a crucial assumption that cannot be discarded. The approach to definitions of cognitive semantics assumed in this article preserves bipolarity in its description of a relation between autoepistemic membership statements and the mental references assigned to these statements.

Secondly, models of reference for particular cases of cognitive semantics are, within our approach, constructed using a variety of distinct tools for representation, processing, and capturing the details of the empirical experience that is assigned to particular classes of the above-mentioned autoepistemic modal statements and stored within the agent. For example, during the design of a cognitive semantics we simultaneously refer to chosen aspects of organization, a degree of processing and completeness of accumulated experience, while accepting as fact that the reference and its correlation with the symbol (in the sense assumed within the semiotic triangle) are very complex in its nature. Such an approach relates not only to a diverse theoretical apparatus devised by the cognitive grammar [29,30], but also to other proposals such as the Distributed Grammar program [34], which is an integrative theoretical framework aiming to cover the relatively complex structure of natural language utterances in general. More details on the Distributed Grammar program can be found, among others, in [35–39].

Thirdly, the proposed approach to the modeling of semantics of autoepistemic membership statements refers to the research on the symbol grounding problem. The problem was originally stated and discussed in [40,41], where grounding has been defined as a process of formation of a relationship between a language symbol and corresponding mental content within the agent's mind. Grounding has been pointed out as the main constituting element of a language for semantic communication. Grounding defined as such has an obvious correlation with cognitive semantics which can be seen as an explicit model of grounding of a symbol. A relationship between the grounding problem and cognitive semantics has not been deeply analyzed in the literature yet, as the research on the grounding problem (and on a related symbol anchoring problem [42]) mainly focused on simple languages consisting of names and labels and their applications, e.g., in robotics [3–6,43,44]. An interesting outcome of the mentioned research is a finding that underlines a function of social learning in the collective shaping of symbols' meaning in semantic communication. As a consequence, a strategy for learning categories with a prototype proposed later in this

paper ensures that cognitive semantics for languages of names and labels is a result of a socially-realized process of semiosis.

Fourthly, this research needs to fulfill requirements formulated within models of conceptual spaces [45–47] which not only assume an existence, at the mental level, of a universe of internal representations of objects but further equip this universe with cognitive tools allowing the agent to meaningfully experience (and evaluate) a level of difference between said objects. These tools usually assume the practical form of a cognitive distance or a cognitive similarity function. A notion of cognitive distance has been directly applied to the mechanisms used for defining cognitive semantics which are presented further in this paper.

Fifthly, as suggested by a focus in the presented research on the case of cognitive semantics for beliefs encoded in semantic communication using symbolic structures of a form $\Sigma(x \in c)$ (assuming their commonsense interpretation), there is an important correlation between the methods and computational models presented in this paper and a wide range of research projects dedicated to the modeling and analysis of so-called modal operators of belief and knowledge. Among a variety of heterogeneous proposals, there is a group of results, mostly of theoretical nature: the original proposal of Hintikka presented in 1962 in his influential work [48]; implementation-wise ideas about autoepistemic logic (with autoepistemic logic understood as in [49–51]); more practice-oriented approaches given as theories, architectures, and programming languages following the so-called BDI (Belief-Desire-Intention) approach to agency, founded mainly on the possible worlds semantics [52]. A range and characteristic of BDI-related approaches have been extensively reported, among others, in [53–59]. Without further deliberation, we only want to state that the methods and models of belief representation proposed in this article, as well as in our previous works, e.g., [18–20,26], differ from a typical BDI approach mostly for two main reasons:

- We abstain from the goal of building a belief calculus (which typically serves as a deductive mechanism for building theorems expressed in a modal language for representing beliefs), and instead focus on the very fundamental level, that is, on a semantic mechanism for the construction of language representations which directly relies on existing mental structures of the agent (treated as a subject of knowledge), while maintaining a requirement of commonsense consistency of the final communicative behavior of the agent.
- We follow a tendency of multiple natural languages and maintain three distinct levels (states) of confidence regarding an agent's beliefs. It extends other approaches which typically assume the following two states of belief confidence: a state of full confidence in regard to the consistency of beliefs with reality and a state of non-exclusive beliefs—allowing complementary beliefs to be present within the agent's mind.

3. The Intuitive Meaning and Syntax of Autoepistemic Membership Statements

As was already mentioned above, we investigate and design a set of fundamental tools for the management of these cognitive processes of an artificial agent which are responsible for a generation of autoepistemic membership statements of a form $\Sigma(x \in c)$, where Σ represents one of the labels of three possible levels of belief confidence, x refers to a particular object (in the simplest case an atom object located in the external world), and c refers to a category with a prototype. However, it was also suggested above that the string of signs $\Sigma(x \in c)$ (being a formula of a semantic communication language) will be treated in our approach as an external language representation of an agent's beliefs as we let ourselves, to some extent and intentionally, abstract from a syntactic form of a particular natural language. The simplification is intentional because our goal is primarily to focus attention on the main logical and semantic elements constituting the cognitive semantics of statements, which do not include, for example, the mechanism of attention or cognitive elements responsible for prosodic effects. The latter mechanisms are responsible for variants of the way in which the agent's belief of the same content can be presented in individual

natural languages. It is also worth recalling that their integration with other cognitive mechanisms participating in the construction of external linguistic representations is the focus of research carried out under the so-called Distributed Grammar program [34–39].

Taking into account the above simplification and using the style of presentation in which similar cognitive states related to beliefs were originally described by Hintikka [48], the range of references (agent's cognitive states) being vertices of the instantiations of semiotic triangles built for the formula $\Sigma(x \in c)$ can be characterized by the following list of extended utterances in English:

- "According to all my collected experience, [*I am certain that* | *I am sure that* | *I know that*] object *o* belongs to category *c*."
- "According to all my collected experience, [*I am certain that* | *I am sure that* | *I know that*] object *o* does not belong to category *c*."

both produced when the agent fully believes in object *o* belonging to category *c*; and

- "According to all my collected experience, *I believe that* object *o* belongs to category *c*."
- "According to all my collected experience, *I believe that* object *o* does not belong to category *c*."

both produced when the agent believes in object *o* belonging to category *c* and this belief is not full but substantially intensive; and finally

- "According to all my collected experience, *I find it possible that* object *o* belongs to category *c*."
- "According to all my collected experience, *I find it possible that* object *o* does not belong to category *c*."

both produced when the agent is in a state of mind in which the above belief is not absent at all but is significantly reduced in intensity.

Obviously, due to their length and apparently extended scope of the communicated mental content (consider the introductory references to empirical experience collected by the uttering agent), the above utterances would probably be treated as a little unusual (even redundant) for everyday natural language, and in everyday speech the following shorter but only slightly less informative messages would be probably used instead of the above long and detailed form of communication:

- "[*I am certain that* | *I am sure that* | *I know that*] *o* is *c*."
- "[*I am certain that* | *I am sure that* | *I know that*] *o* is not *c*."
- "*I believe* that *o* is *c*."
- "*I believe* that *o* is not *c*."
- "*It is possible* that *o* is *c*."
- "*It is possible* that *o* is not *c*."

An analogous list of extended utterances in Polish can take the following form:

- "Ze względu na całość mojego doświadczenia, [*jestem pewien, że* | *wiem, że*] obiekt *o* należy do kategorii *c*."
- "Ze względu na całość mojego doświadczenia, [*jestem pewien, że* | *wiem, że*] obiekt *o* nie należy do kategorii *c*."
- "Ze względu na całość mojego doświadczenia, *sądzę, że* obiekt *o* należy do kategorii *c*."
- "Ze względu na całość mojego doświadczenia, *sądzę, że* obiekt *o* nie należy do kategorii *c*."
- "Ze względu na całość mojego doświadczenia, *możliwe, że* obiekt *o* należy do kategorii *c*."
- "Ze względu na całość mojego doświadczenia, *możliwe, że* obiekt *o* nie należy do kategorii *c*."

and their shorter (non-redundant) counterparts can be stated as follows:

- "[*Jestem pewien, że* | *Wiem, że*] *o* jest *c*."

- "[*Jestem pewien, że* | *Wiem, że*] *o* nie jest *c*."
- "*Sądzę, że o* jest *c*."
- "*Sądzę, że o* nie jest *c*."
- "*Możliwe, że o* jest *c*."
- "*Możliwe, że o* nie jest *c*."

We treat the above-mentioned non-negative utterances as equivalents (instantiations), among others, of the general formula $\Sigma(x \in c)$, expressed in a controlled natural language, additionally distinguishing three cases $Know(x \in c)$, $Bel(x \in c)$, $Pos(x \in c)$ for labeling of the three possible levels of an agent's belief confidence described in the introduction. Obviously, an analogous remark, but with formula $\Sigma(x \notin c)$, should be applied to these examples.

Keeping in mind the above characterization of intuitive (commonsense-grounded) meaning of autoepistemic membership statements, for technical purposes we will now capture the whole language of modal categorization in a more concise way, labeling it by K:

Definition 1. *The alphabet of the modal membership language K and an auxiliary language of atomic membership statements K^N consists of the following elements:*

- *symbols $x \in X$ unequivocally pointing to atomic objects of the agent's external world,*
- *names of categories $c \in C$; being string literals e.g., bird, robin,*
- *symbol $\in$ for the binary relation defined over the set $X \times C$,*
- *symbol $\notin$ for the binary relation defined over the set $X \times C$,*
- *symbols Pos, Bel, Know; being representations of modal autoepistemic operators,*
- *auxiliary symbols '(' and ')'.*

Definition 2. *The auxiliary language of atomic membership statements K^N is given as:*

$$K^N = \{x \in c : x \in X \text{ and } c \in C\} \cup \{x \notin c : x \in X \text{ and } c \in C\}. \tag{3}$$

Definition 3. *The modal membership language K is given as:*

$$K = \{Know(\varphi) : \varphi \in K^N\} \cup \{Bel(\varphi) : \varphi \in K^N\} \cup \{Pos(\varphi) : \varphi \in K^N\}. \tag{4}$$

The assumed intuitive (commonsense-grounded) meaning of members of K^N and K is briefly summarized in Tables 1 and 2. It is worth recalling that the language K was previously considered in [18]. However, it was used there to communicate the belief that the object belongs only to the category without a prototype.

Table 1. Intuitive semantics of non-modal atomic formulas.

Formula	Intuitive Meaning
$x \in c$	Object x belongs to category c.
$x \notin c$	Object x does not belong to category c.

Table 2. Intuitive semantics of modal atomic formulas.

Formula	Intuitive Meaning
$Know(x \in c)$	I know that object x belongs to category c.
$Know(x \notin c)$	I know that object x does not belong to category c.
$Bel(x \in c)$	I believe that object x belongs to category c.
$Bel(x \notin c)$	I believe that object x does not belong to category c.
$Pos(x \in c)$	I find it possible that object x belongs to category c.
$Pos(x \notin c)$	I find it possible that object x does not belong to category c.

4. Model of the Agent

4.1. Internal Architecture of the Agent

R&D goals set in the introduction require a formulation of a dedicated internal architecture of an artificial agent allowing for an effective execution of key cognitive processes responsible for a formation (often through learning in a social setting) and management of categories with prototypes and for further usage of cognitive semantics based on these categories during communication acts. In order to realize these two main capabilities within an agent, its internal architecture needs to contain relevant internal structures. Figure 2 outlines such an architecture which contains, among others, the following set of interlinked components:

- basic mental space,
- repository of categories,
- repository of cognitive semantics,
- working memory,
- repository of episodes,

which are directly related to the investigated phenomena, that is, to a set of cognitive functions realized by:

- an interface for perception of the external environment,
- a module which governs the choice of a symbol of a semantic communication (with a focus on autoepistemic membership statements),
- natural language generation interface.

Many of these processes must run concurrently in an agent. Therefore, the final implementation is bound to have to pay strong attention to the way in which they are dynamically interlinked within the agent's artificial mind to ensure required dialogue capabilities [60].

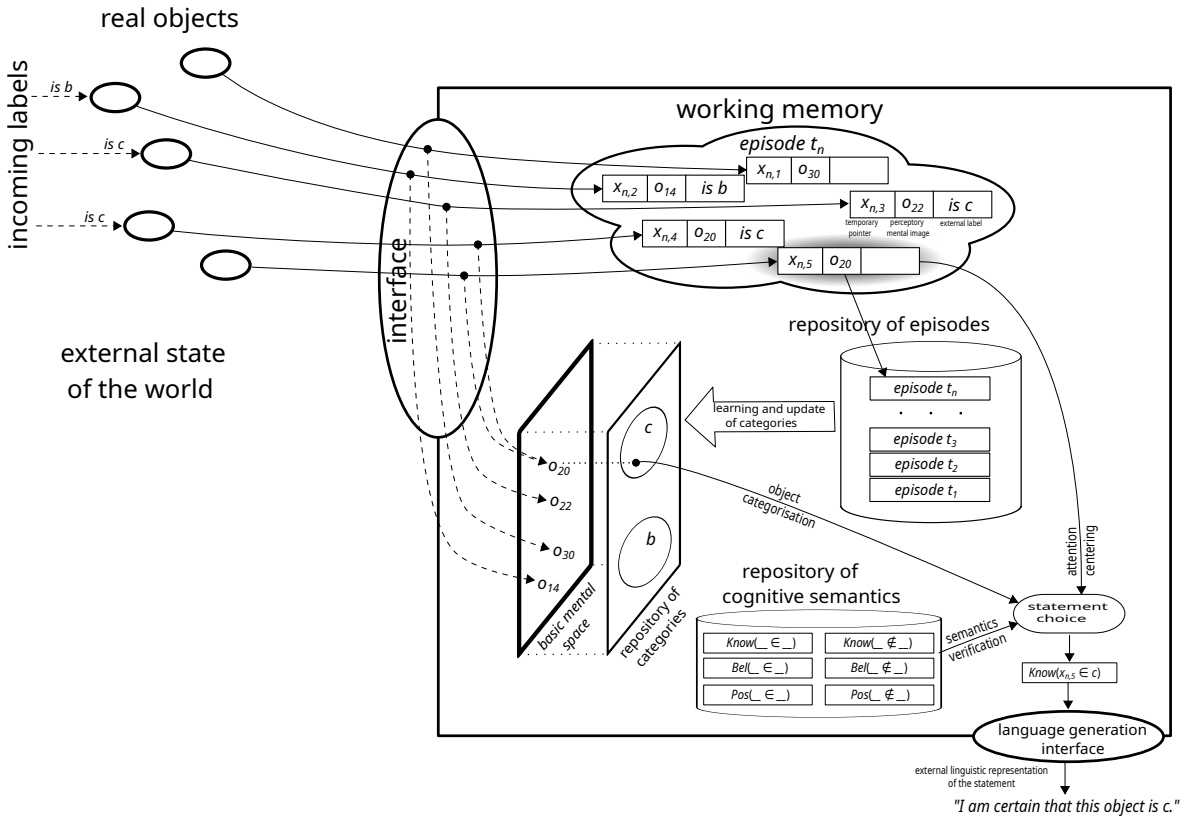

Figure 2. Architecture of the agent.

4.2. Model and Role of Basic Mental Space

From a cognitive perspective, the basic mental space is a key component of the architecture as it represents a fundamental cognitive competence of an agent of being able to distinguish objects of a particular universe. The elements of the basic mental space exhibit the nature of atomic (indivisible) entities, whose state, behavior, and relations with other elements can become an object of particular beliefs of an agent.

From this article's goals' standpoint, there are two crucial functionalities of the basic mental space. At first, elements of the basic mental space allow an agent to order (organize and conceptualize) an empirical material obtained through direct observation of the external environment. Contents of these observations is restricted to objects whose existence is entailed by the structure of the basic mental space. Elements not expressible in the basic mental space, provided they exist, are in this research treated, from the agent's point of view, as cognitively ungraspable, and hence completely omitted.

At second, the basic mental space serves as a fundamental layer over which categories (in particular, investigated categories with a prototype) can be defined. But more importantly, we claim that each conceptual category has to be always defined in relation to a respective mental space. It entails that the basic mental space adopted in the presented architecture of an artificial agent, after equipping it with tools for a formal (numerical) expression of a degree of similarity (or distance) between any pair of objects from the mental space, corresponds directly to the aforementioned theory of the conceptual spaces [45–47].

We adopt the following definition of the basic mental space:

Definition 4. *Universe of mental representations of distinguishable objects O is defined as follows:*

$$O = \prod_{a \in A} V_a = \{o_1, \ldots, o_M\}, \tag{5}$$

where

- *A denotes a finite set of attributes,*
- *for each $a \in A$, V_a denotes a domain of attribute a.*

We assume that the domains V_a of attributes $a \in A$ depend on the nature of objects considered by the agent and due to the range of practical applications we allow, they may contain values of the following type:

- binary—e.g., $\{0, 1\}$, $\{yes, no\}$,
- nominal—e.g., $\{large, medium, small\}$, $\{red, green, blue, black\}$,
- numeric—e.g., set of natural numbers, set of real numbers.

Obviously, despite being a phenomenological whole, from the agent's point of view each element $o \in O$ is cognitively decomposable into specific values of attributes, which are still integral components of this object. It is worth noting that such a decomposition is integrally related to the process of generating statements about a specific feature of an object and, in a broader theoretical perspective, refers to one type of the so-called predication discussed in formal logic, cognitive science and, above all, in linguistics (e.g., [37,38]).

The way in which the basic mental space participates in building a mental representation of an agent's observations of external worlds, as well as in obtaining and representing a model of a category with a prototype, is presented in the following parts of this article.

4.3. Repository of Categories with Prototype

As can be seen in Figure 2, it is assumed that models of all categories known to the agent are stored in a dedicated repository. The repository is an important part of the internal representation of the so-called embodied ontology, representing how the agent perceives the conceptual order of the external world. Another basic element of such an ontology is the basic mental space introduced above.

Each category with a prototype is built in relation to the accessible mental basic space O. In our approach, we try to reflect as many structural and functional features as possible, which have been provided for categories in the so-called standard version of the prototype semantics [9]. In particular, each category is assigned a specialized object $o_c^\star \in O$ distinguished from all other objects in this category, called a prototype of the category. It is perceived by the agent as being the most representative and in some sense "central" element of the category.

An agent's attitude towards a membership of an object to a given category should be directly related to the degree of that object's similarity (or distance) to the prototype. Such an approach translates to a requirement that the agent should be equipped with a cognitive ability allowing for comparison of any object $o \in O$ with a prototype $o_c^\star$ of any category c. Additionally, this ability enables the agent to order objects according to their similarity (or distance) to the prototype. Naturally, the most similar (closest to the prototype) objects are considered as belonging to the category. On the other hand, the least similar (farthest from the prototype) objects are considered as not belonging to the category. This is going to be determined within the agent by a set of ranges assigned to the category by the agent itself during the process of category learning. Consistent with the assumption that the boundary of a category may be fuzzy, our model reflects it by using ranges in order to introduce an area of the mental basic space in which a membership of an object to a category is considered uncertain (understood here as *partial* rather than as *probable*) by the agent. These ideas lead to the formulation of category's model consisting of the following:

- a prototype of a category—a mental representation of a prototypical object, considered the very center of the category,
- a core of a category—an area of a basic universe of mental representations naturally assigned by an agent to the category; for natural reasons, the core should surround the prototype,
- a boundary of a category—an area of a basic universe of mental representations where an applicability of the category is questionable,
- an outer of a category—an area of a basic universe of mental representations where the concept is not applicable.

Assuming the distance-based approach, perhaps in the simplest but still effective way the category can formally be defined as a triplet:

$$\langle o_c^\star, \tau_c^+, \tau_c^- \rangle, \tag{6}$$

where:

- $o_c^\star$ is a prototype of c,
- τ_c^+ is a radius defining *Core* of c,
- τ_c^- is a radius defining *Boundary* of c,
- and the following condition holds:

$$\tau_c^+ < \tau_c^-. \tag{7}$$

Figure 3 graphically presents a deliberately simplified model of a category with a prototype using a visual notion of the cognitive distance (reflected as graphical distance between points on the plane). The category is defined over the basic mental space $O = \{o_1, \ldots, o_8\}$. The object o_7 is indicated as the prototype of the category. Objects o_1 and o_7 are located in the core of the category and therefore are both perceived as belonging to the category c, although object o_1 is not assessed by the agent as fully conforming to the prototype o_7. Objects o_2, o_5 and o_8 are considered by the agent not only as different from the prototype of c, but also as objects with their membership to the category being uncertain. While objects o_3, o_4, and o_6 are treated as not belonging to the category c.

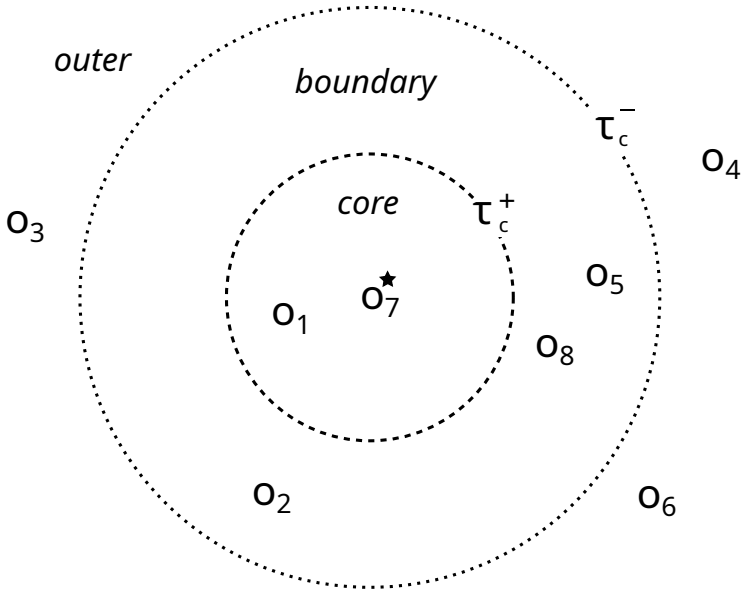

Figure 3. Schematic picture of three regions of the model of a category.

This interpretation shows in what way the cognitive experience related to our model of the category with a prototype should be treated as naturally multidimensional, in particular, when it comes to the relationship of individual objects of the basic mental universe O with the category. Such a relation includes both the aspect related to objects' location relative to the prototype, as well as the aspect of the degree of objects' membership to the fuzzy boundary of the category. Both aspects need to be adequately addressed in any implementation of an agent to ensure its proper linguistic behavior.

4.4. Repository of Episodes

Basic experience related to the agent's knowledge about the external world is collected in an internal repository of the so-called episodes. The knowledge stored in the episodes relates to two dimensions of the agent's cognition: direct perception of the state of objects of an external world in which the agent operates, and direct recognition of some semantic language messages produced by other agents (all together treated as a collective teacher) in relation to particular objects and representing statements about membership of these objects to some categories. It means that the basic experience stored in the episodes covers some aspects of physically and socially grounded experience. As a consequence, the functionality of the interface assumed in the proposed architecture (see Figure 2) is required to cover, among others, the following two groups of interrelated actions: at first, recognizing and internally representing actual states of external objects; at second, recognizing messages of a semantic language of communication and then building internal representations of their meaning.

The ability to cope with the social aspects of the external world is a prerequisite for an agent's competence in any semantic communication, which must always be grounded in a fixed social context, as well as learned in such a context. In our model, we assume two kinds of messages incoming from externally located agents (as a group interpreted as a collective teacher) with each represented by an element from set L containing the following labels:

- *is-c*, called a positive label indicating the category c, used to represent an external agent's certainty that a labeled object belongs to the category c,
- *not-c*, called a negative label indicating the category c, used to represent an external agent's certainty that a labeled object does not belong to the category c.

An important feature of the repository of episodes is that each episode is related to a particular time point, which means that in the model a specific competence of the agent in the processing of the temporal aspects of knowledge representation is assumed. Namely,

each episode is treated as being related to a particular time point $t \in T = \{t_0, t_1, t_2, \ldots\}$, where for each $i = 1, 2, \ldots$, time point t_i is interpreted as an immediate predecessor of t_{i+1} in linear temporal order. For formal purposes, we will also use the following relation of temporal precedence: $\leq^{TM} = \{(t_i, t_j) : t_i, t_j \in T \wedge i \leq j\}$.

Finally, the following definition of an episode is adopted:

Definition 5. *The episode, interpreted as an internal model of a particular observed state of an agent's external world along with recognized linguistic labels assigned in this state to some objects, is given as a tuple* $Episode(t) = \langle X_t, A, V, L_t, percept, label \rangle$. *The tuple is related by the agent to a particular time point t and its elements are interpreted as follows:*

- X_t *denotes a finite set of individual representations of objects recognized in the state of environment,*
- A *denotes a set of attributes assigned to objects (according to the adopted definition of the basic mental space),*
- $V = \bigcup_{a \in A} V_a$, *where* V_a *denotes a domain of a particular attribute a,*
- L_t *denotes a finite set of adopted labels,*
- $percept_t$ *denotes a total function* $X_t \times A \longrightarrow V$, *such that for all* $x \in X_t$ *and* $a \in A$ $percept_t(x, a) \in V_a$ *and represents a value of attribute a observed by the agent for object x,*
- $label_t$ *denotes a (not-necessarily total) function* $X_t \longrightarrow \Pi(L)$, *such that for all* $x \in X_t$ $label_t(x) \subseteq L$ *contains all the labels recognized by the agent as being assigned to object x in the observed state of the external world.*

From the formal point of view, the concept of an episode just introduced can be treated as an extension of the classic definition of a single-valued and complete information system [61,62]. The part of an episode which is equivalent to a complete information system is used by the agent to represent the result of a particular observation of a certain state of the external world. The extension is used as a collection of other agents' beliefs about the membership of objects cognitively distinguished in the observed state of the external world to specific conceptual categories. Recognition of other agents' beliefs is based on observed occurrences of the labels introduced above *is-c* or *not-c*.

It is important to note that individual episodes are constructed by the agent in working memory and then are used to update the main repository of episodes (see Figure 2). In order to distinguish individual states of the overall empirical knowledge base at particular time points (given by all collected episodes), the following symbol is additionally introduced:

Definition 6. *At each time point* $t \in T$, *the state of empirical knowledge about the external world is defined by a temporal collection of episodes given as follows:*

$$Episodes(t) = \{Episode(t_n) : t_n \in T \quad and \quad t_n \leq^{TM} t\}. \tag{8}$$

The following example illustrates the way in which the above concepts are used to represent basic knowledge about the external world, stored internally in the agent. It is based on a deliberately simplified knowledge base, defined under the influence of the work [12] where the main considered objects were *birds* described by a mutual set of attributes (including attributes: *having beak or bill*, *having wings*, *dominant color*, further in the example denoted by a_1, a_2 and a_3, respectively.

Example 1. *The Table 3 shows the internal knowledge base built by an agent up to the time point* t_2. *It is represented by the collection* $Episodes(t_2) = \{Episode(t_1), Episode(t_2)\}$.

In the first recognized state of the external world, the agent could recognize three objects, so therefore the following three-element set of objects $X_{t_1} = \{x_{1,1}, x_{1,2}, x_{1,3}\}$ *was created as a component of* $Episode(t_1)$. *Meanwhile, in the second state of the world, only the following two-element set* $X_{t_2} = \{x_{2,1}, x_{2,2}\}$ *was created as a component of* $Episode(t_2)$.

The minimum basic mental space O that would allow for the agent to distinguish the objects listed above and described in Table 3 would have to be defined over the set of attributes given as follows: $A = \{a_1, a_2, a_3\}, V_{a_1} = \{yes, no\}, V_{a_2} = \{yes, no\}, V_{a_3} = \{black, white\}$. *As a consequence, the related cognitive competence of the agent in recognizing objects at all would be constrained to the eight-element set* $O = V_{a_1} \times V_{a_2} \times V_{a_3}$. *Provided that the minimum basic mental space is given as in Table 4, the following can be additionally said: objects* $x_{1,1}$ *and* $x_{2,1}$ *are treated by our agent as mental representations of two externally observed instantiations (realizations) of mentally cognizable objects* $o_1 \in O$. *A similar interpretation applies to objects* $x_{1,2}$, $x_{1,3}$ *and basic mental object* $o_2 \in O$, *as well as object* $x_{2,2}$ *and basic mental object* $o_8 \in O$.

By carrying out the observations of both distinguished states of the world, the agent also acquired socially grounded knowledge about how other agents (interpreted as a collective teacher) classified objects from sets X_{t_1} *and* X_{t_2} *into categories known to them. Namely, in relation to the state marked by time point* t_1, *the agent found out that the collective teacher marked all objects* $x_{1,1}$, $x_{1,2}$, *and* $x_{1,3}$ *with language means represented by a positive label is-bird, this way expressing socially grounded belief that all of them belong to the category referred to by bird. The knowledge may be formalized by function* $label_{t_1} : X_{t_1} \longrightarrow \Pi(\{is\text{-}bird, not\text{-}bird, is\text{-}mammal, not\text{-}mammal\})$, *with values given as in Table 3.*

Similar knowledge concerning the second recognized state concerned only object $x_{2,2}$. *The collective teacher marked with two labels (negative not-bird and positive is-mammal) what in our agent's cognitive perspective was interpreted as representation of belief that the object* $x_{2,2}$ *does not fall into the category bird, but does belong to the category mammal. In* $Episode(t_2)$ *object* $x_{2,1}$, *however, was not accompanied by representation of any linguistic messages regarding the object's belonging to any category provided by external agents. This knowledge of external beliefs may be represented by the function* $label_{t_2} : X_{t_2} \longrightarrow \Pi(\{is\text{-}bird, not\text{-}bird, is\text{-}mammal, not\text{-}mammal\})$, *again, with values given as in Table 3.*

Table 3. Two exemplary episodes.

Episode	Object	a_1	a_2	a_3	Labels
Episode (t_1)	$x_{1,1}$	yes	yes	black	is-bird
	$x_{1,2}$	yes	yes	white	is-bird
	$x_{1,3}$	yes	yes	white	is-bird
Episode (t_2)	$x_{2,1}$	yes	yes	black	
	$x_{2,2}$	no	no	white	not-bird, is-mammal

Table 4. Example basic mental space.

Basic Mental Object	a_1	a_2	a_3
o_1	yes	yes	black
o_2	yes	yes	white
o_3	yes	no	black
o_4	yes	no	white
o_5	no	yes	black
o_6	no	yes	white
o_7	no	no	black
o_8	no	no	white

4.5. Considered Functions of Episodes

Potential functions of individual episodes within the proposed agent's architecture can be diverse. The following two are particularly interesting in relation to the processing of autoepistemic membership statements:

- being an internal entity contributing to the creation and update of the models of categories with prototypes,
- being a time point-related context for the production of membership statements.

The first of them is fundamental as a function that enables the agent to carry out an autonomous process of learning and updating socially valid models of categories with prototypes on the basis of examples of objects' categorization obtained from individual acts of semantic communication with other agents (that are all treated in this work as a single collective teacher).

This fairly obvious possible relationship between the content of an individual episode and category learning is depicted in Figure 4 related to Example 1.

The main connection between the episode and the learning of categories are the elements of basic mental space O. Namely, in accordance with the assumption regarding the agent's competence in mental conceptualization of objects in general, each element x in any set X_t must and does have one and exactly one corresponding element $o \in O$. Such an element o plays the role of a conceptual pattern of x. In Example 1, this type of relationship was represented by pairs: $(x_{1,1}, o_1)$, $(x_{1,2}, o_2)$, $(x_{1,3}, o_2)$, $(x_{2,1}, o_1)$, and $(x_{2,2}, o_8)$. On the other hand, each object $o \in O$ is an element of the universe O over which models of particular conceptual categories are defined, including categories *bird* and *mammal*, both stored in the example repository of categories. For this reason, gaining knowledge about relationships between models of external objects represented by elements in sets X and labels indicating specific conceptual categories, assigned by the collective teacher, allows for the agent to first establish, and then update, the internal model of the current social consensus regarding the meanings of the categories. In particular, such knowledge can and should influence the shape of the core, boundary, and outer components of relevant categories.

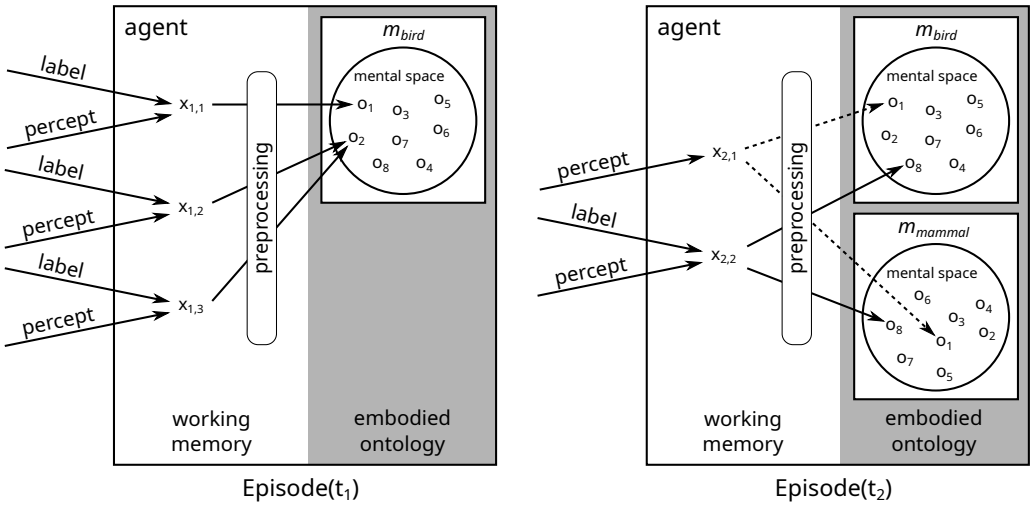

Figure 4. Preprocessing of two exemplary episodes.

The second one of the above two distinguished functions, i.e., being a time point-related context for the production of autoepistemic membership statements, is strictly related to the range of semantic meanings of the statements chosen to be considered in this article. Namely, it was assumed in Definition 1 that each symbol $x \in X$, when it is used as a part of an individual autoepistemic membership statement, is an unambiguous pointer to a particular atomic object of the agent's external world. It means that the assumed commonsense and pragmatic interpretation of $x_{i,j} \in X_t$, $_{i,j \in \{1,2,\ldots\}}$ refers unequivocally to Definition 5 where the result of each cognitive experience resulting in the mental distinction of a particular external object within an observation covering a particular state of the external world is represented by a dedicated pointer that corresponds to that object. Such a pointer is explicitly included in set X_t which is a basic ontological component of the episode internally representing that state of the external world.

This remark explains why each episode, apart from its basic function of representing a certain state of the world, should also be treated as a complex cognitive structure mediating between any description of a state of the external world expressed by the agent in a chosen

natural or semi-natural language, and the reality to which the description applies. Taking a slightly different perspective, it can also be said that episodes are elements of representation of the agent's inherently subjective beliefs about external and observed individual states of the environment.

The following Example 2, again deliberately simplified, extends Example 1 and shows how the second function of episodes should be adopted within the agent's architecture during natural language production.

Example 2. *Let us consider the agent originally introduced in Example 1 and assume that from the agent's point of view time point t_3 labels a current state of the external world. Let us also assume that the repository of category models available for the agent contains two models of categories bird and mammal (see Figure 5). While observing the current state of the external world, the agent has built $Episode(t_3)$ as given in Table 5. It means that at time point t_3 which is internally treated as related to the current state of affairs, the agent at first perceives two (and only two) external objects $x_{3,1}$ and $x_{3,2}$, as well as how at second there is no label perceived by this agent related to any of the recognized objects. The latter can be formally described by $label_{t_3}(x_{3,1}) = label_{t_3}(x_{3,2}) = \emptyset$. Obviously, the state of repository related to the current time point t_3 is given as $Episodes\,(t_3) = Episodes(t_2) \cup \{Episode\,(t_3)\}$.*

Table 5. The third episode.

Episode	Object	a_1	a_2	a_3	Labels
$Episode\,(t_3)$	$x_{3,1}$	yes	yes	black	
	$x_{3,2}$	no	no	white	

Let us now recall that the intuitive (pragmatical) interpretation of the autoepistemic membership statements of languages K^N and K, specified in Definitions 1 and 2 (provided that $X = X_{t_3}$), assumes that the statements are spoken in the present grammatical tense. This applies both to autoepistemic extensions represented by operators $\{Pos, Bel, Know\}$, as well as to their arguments belonging to language K^N. It means that the content of beliefs communicated by statements in K is always captured by this episode which the agent treats as a representation of a current state of the external world. Obviously, the mapping of statements of K onto a "current" episode has a conventional (consensual) character and in the case of natural languages is the result from a related socially realized process of semiosis. Therefore, what is included in the episode experienced by the agent as a representation of the current state of the world, determines (and in this sense constraints) the range of linguistic representations generated. Since in the example we assign the role of representing the current state of the external world to $Episode(t_3)$, the range of statements that can be used as potential representations of the agent's own beliefs about the current state of features of external objects is given by the following set:

$$K_{/t_3} = \{Know(x_{3,1} \in bird), Know(x_{3,1} \notin bird), Bel(x_{3,1} \in bird), Bel(x_{3,1} \notin bird),$$
$$Pos(x_{3,1} \in bird), Pos(x_{3,1} \notin bird), Know(x_{3,2} \in bird), Know(x_{3,2} \notin bird), Bel(x_{3,2} \in bird),$$
$$Bel(x_{3,2} \notin bird), Pos(x_{3,2} \in bird), Pos(x_{3,2} \notin bird), Know(x_{3,1} \in mammal),$$
$$Know(x_{3,1} \notin mammal), Bel(x_{3,1} \in mammal), Bel(x_{3,1} \notin mammal), Pos(x_{3,1} \in mammal),$$
$$Pos(x_{3,1} \notin mammal), Know(x_{3,2} \in mammal), Know(x_{3,2} \notin mammal), Bel(x_{3,2} \in mammal),$$
$$Bel(x_{3,2} \notin mammal), Pos(x_{3,2} \in mammal), Pos(x_{3,2} \notin mammal)\}$$

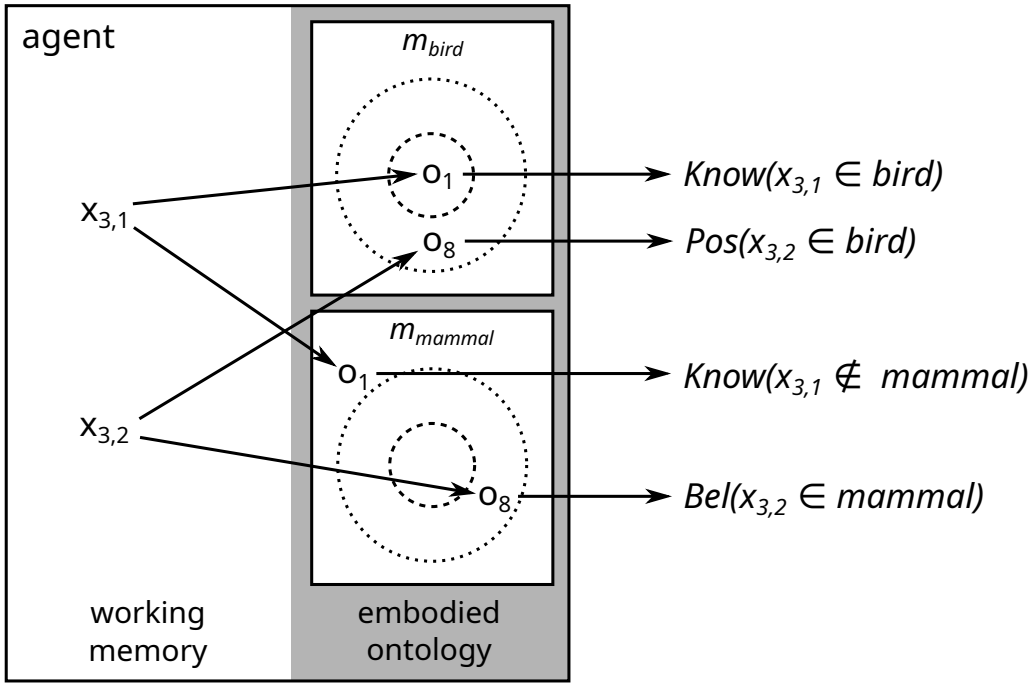

Figure 5. Example categories assumed to be available at a "current" time point t_3 and related grounding of autoepistemic membership statements.

In relation to Example 2, the following two supplementary notes are worth being made at this point in the presentation.

At first, it should now be announced that it is the cognitive semantics, reflecting the social consensus of meaning developed for the considered language K, that will determine which subset of the set $K_{/t_3}$ is ultimately used as an adequate linguistic representation of beliefs about the current state of affairs adopted by the agent. Figure 5 shows a possible example choice and some of the main cognitive components involved in making it. Firstly, the objects $x_{3,1}$ and $x_{3,2}$ are mapped onto the universe O in order to then be projected onto models of particular categories. Then, cognitive semantics, taking into account some additional numerical characteristics describing the latter mapping and specified explicitly in further definitions of cognitive semantics, determines the final range of statements representing the considered state of beliefs. As is assumed in Figure 5, such a possible set of relevant statements might be $\{Know\,(x_{3,1} \in bird), Pos\,(x_{3,2} \in bird),\}$ $Know\,(x_{3,1} \notin mammal), \{Bel\,(x_{3,2} \in mammal)\}$, although the range of possible choices is not limited to the given case. The reason will be explained later in this article.

Secondly, in this article we consider a simplified version of the possible pragmatic context within which the considered cases of autoepistemic membership statements are generated. Namely, we establish that the agent's "centered" attention covers the entire content of the episode representing the current state of the external world and the agent's goal is to select a complete set of statements representing beliefs related to all cognitively available conceptual categories and to all objects distinguished in this and only this episode. However, a goal constructed in this way rarely occurs in practical situations. Indeed, typical uses of statements being under consideration are usually part of a dialogue involving more than one agent and in such a dialogue the production of a specific statement by a specific agent is usually a response to a query addressed to this agent. Possible examples of such questions are *Does this object belong to category c?*, *What is* $x_{3,1}$*?*, etc. The introduction of such or similar detailed and specific pragmatic contexts can and in fact leads to a significant expansion of the range of symbolic representations used. For example, in the case of the English language, it is enough to note that external (symbolic) references to the object $x_{3,1}$ appearing in $Episode\,(t_3)$, with the additional assumption of shared attention of agents participating in a specific dialogue, may take the form: *this object*, *that object*, *it*, etc., instead

of *object* $x_{3,1}$ adopted in this article. Moreover, to be complete, such a list should also include extensive descriptive pointers to the object $x_{3,1}$, e.g., *the object that is located here/there/next to*, etc. The architecture we have considered has been deliberately simplified and does not include conceptual tools that, at the level of the agent's internal knowledge bases, would allow for representing more detailed and specific contexts of the potential use of autoepistemic membership statements and, consequently, variant cognitive semantics related to beliefs about the belonging of currently observed objects to categories with prototypes. However, it is highly probable that changes to the proposed architecture that would enable the representation of more detailed and more specialized statement generation contexts would consist solely in extending the proposed architecture rather than eliminating or significantly changing the form of its already proposed elements.

5. The Strategy of Initial Learning of Categories with Prototypes

The way in which the proposed agent autonomously acquires and then updates cognitively accessible categories with a prototype is another element of the complex set of cognitive processes that are implemented within the architecture given in Figure 2.

5.1. The Input Learning Data

Following a commonsense claim stated by Dennett [63] that "*exposure to x—that is, sensory confrontation with x over a suitable period of time—is the normally sufficient condition for knowing (or having true beliefs) about x*", we have decided to perform a process of category learning based on a series of such exposures, namely the agent's interactions with the external world, each represented by a particular episode and, among others, containing both positive (*c* confirmed—labeled as *is-c*) and negative (*c* denied—labeled as *not-c*) learning examples, as was introduced above. Therefore, in forthcoming parts of this article, the input collection of the learning examples to be used in acquiring a specific category *c* will be a multiset denoted by Exp_c, defined over the universe $O \times \{is\text{-}c, not\text{-}c\}$, and represented as in the following formula:

$$Exp_c = \{k^+_{o,c}(o, is\text{-}c), k^-_{o,c}(o, not\text{-}c)\}_{o \in O} \tag{9}$$

where:

- $k^+_{o,c}$ is a non-negative integer denoting a number of learning examples where *o* has been labeled by *is-c*,
- $k^-_{o,c}$ is a non-negative integer denoting a number of learning examples where *o* has been labeled by *not-c*.

A particular method by which the content of Exp_c can actually be extracted from an available collection of episodes depends always on practical context and rational choice criteria for learning data.

Example 3. *Let us assume that at a time point* t_3*, the agent with cognitive competence assumed as in Examples 1 and 2 is equipped with no model of a category with a prototype. However, there is some potentially learning content available in the repository* $Episodes(t_3)$ *consisting of positive and negative labels communicating membership of at least some observed objects to two categories bird and mammal. Although the list of these labels is only illustrative, let us assume that the agent launches a strategy dedicated to generate the first version of models for the categories bird and mammal, which involves generating these models based on all available training materials. An alternative version of such a strategy could omit, for example, the oldest labels (which is impossible in this deliberately simplified case). In such a situation, the following formulae are adopted:*

$$k^+_{o,c} = card\Big(\{(x, is\text{-}c) : \exists t_p \leq t \; \exists x \in X_{t_p} \; (x = o \wedge is\text{-}c \in label_{t_p}(x))\}\Big), \tag{10}$$

$$k^-_{o,c} = card\Big(\{(x, not\text{-}c) : \exists t_p \leq t \; \exists x \in X_{t_p} \; (x = o \wedge not\text{-}c \in label_{t_p}(x))\}\Big) \tag{11}$$

and the resulting learning experience, used to create (to learn) initial models of categories bird and mammal, is aggregated in the form of the following two multisets, respectively: $Exp_{bird} = \{1(o_1, is\text{-}bird), 2(o_2, is\text{-}bird), 1(o_8, not\text{-}bird)\}$ *and* $Exp_{mammal} = \{1(o_8, is\text{-}mammal)\}$*. For simplicity, pairs with zero counters are omitted.*

5.2. Outline of the Category Learning Steps

This section builds upon a strategy for learning categories with prototypes, which was in a brief manner presented in [64] to show how categories can be computationally and relatively consistently with intuition extracted from collected learning experience. The agent starts with the learning experience related to a particular category c aggregated into the above introduced multiset Exp_c (see Formula (9)). It intends to derive a model of a category with a prototype c, that is, the category's prototype, core, and boundary. Following the model of category with prototype introduced in Section 4.3, the algorithm goals are, in particular, to find a prototype $o_c^\star$ and to determine two related radii τ_c^+ and τ_c^-.

The basic ideas behind the algorithm are as follows: at first, to use an adopted basic mental space enriched with a specific distance or similarity function f (see Definition 4); at second, to determine candidates for prototype; at third, to analyze all the candidates according to some criteria; and at the end, to choose the prototype, if it exists. It is generally expected that the prototype should be surrounded by a core of the category with the currently largest possible radius, and contain a substantially large part of the supporting (positive) examples of the learning data. In a case where no candidates fulfill such requirements, the procedure should fail which means that the current experience does not support building a model of category c as a category with a prototype.

The algorithm iterates over objects $o \in O$ occurring (mentioned) in the learning experience. Therefore, to make further presentation more convenient, some additional notions (subsets) are derived from Exp_c:

- $E_c^+(Exp_c) = \{o : o \in O \wedge k_{o,c}^+ > 0\}$ denotes a set of objects $o \in O$ supported by at least one positive example of learning experience given in Exp_c,
- $E_c^-(Exp_c) = \{o : o \in O \wedge k_{o,c}^- > 0\}$ denotes a set of objects $o \in O$ supported by at least one negative example of learning experience given in Exp_c,
- $\hat{E}_c^+(Exp_c) = \{k_{o,c}^+(o, is\text{-}c) \in Exp_c\}$ denotes a multiset aggregating positive part (labeled by *is-c*) of learning experience given in Exp_c.

An idea behind the method for category formation [64] is to start with a set of suitable candidates for a prototype of a category for which the agent is trying to build an internal model. Omitting the whole philosophical discussion related to whether a prototype should be a reflection of a particular existing object the agent has ever observed in the past, or whether hypothetical objects from the whole mental space are also allowed, we skip to the conclusion by saying that the prototype should be, in a way, a central element of a multiset containing learning experience positively supporting a particular category, namely a multiset $\hat{E}_c^+(Exp_c)$.

An algorithm presented in the next section assumes a particular set of candidates for prototype as being extracted using a function *extractCandidates*. In a majority of practical contexts, candidates are chosen based on certain optimization criteria. When the basic mental space is equipped with a distance function f, usual cases of candidates o_c for prototype $o_c^\star$ are given as follows:

- the most common element of $\hat{E}_c^+(Exp_c)$—an honorable mention rather than a serious option—while it is the easiest one to calculate, it ignores distance-based relations between elements of $\hat{E}_c^+(Exp_c)$. It can be calculated as follows:

$$o_c = \arg\max_{o \in O} k_{o,c}^+, \tag{12}$$

- medoid of $\hat{E}_c^+(Exp_c)$—a central element of $\hat{E}_c^+(Exp_c)$ chosen from among objects experienced by an agent during its learning episodes, calculated as follows:

$$o_c = \arg \min_{o \in O \wedge k_{o,c}^+ > 0} \sum_{x \in O} \left(k_{x,c}^+ \cdot f(o,x) \right), \tag{13}$$

- centroid of $\hat{E}_c^+(Exp_c)$—a central element of $\hat{E}_c^+(Exp_c)$ which allows hypothetical prototypes (not present within relevant learning experience):

$$o_c = \arg \min_{o \in O} \sum_{x \in O} \left(k_{x,c}^+ \cdot f(o,x) \right). \tag{14}$$

Obviously, if f is a similarity function, the Formulas (13) and (14) should be changed by replacing arg min with arg max.

It is important to note that a frequent association of a particular language label to an object o is not enough to set it as a prototype of category c. It is, in particular, required that there is a region around a chosen o (in the sense of a similarity or distance function-based neighborhood) where an association with a particular concept is not ambiguous, that is, it has not been contested within the learning experience by labeling nearby objects by "*not-c*". Therefore, it is important to further verify the validity of extracted candidates. In consequence, all the candidates (usually there is just one but formally there might be more than one element $o \in O$ satisfying the above optimization criteria) are further passed to an algorithm (described in the upcoming section) in order to verify their feasibility and to either discard them or to choose one of them as the final prototype.

5.3. Definition of the Strategy and Extended Computational Example

The category learning steps discussed above are integrated in the form of a strategy which is presented below as Algorithm 1. The input data is given as a learning set $Exp_c \in \hat{\Pi}(T^c)$ and the algorithm consists of the following:

- preparatory steps (lines 2–4) that define a variable $o_c^\star$ used to evaluate a STOP condition of the algorithm, determine sets of objects confirmed as c (the set $E_c^+(Exp_c)$) and rejected as c (the set $E_c^-(Exp_c)$) within the learning experience,
- preparation of a set of candidates (line 5) using *extractCandidates* function,
- a general evaluation loop for candidates (condition in line 6),
- evaluation of distance values and potential radii values, and *Core/Boundary/Outer* for a particular candidate (lines 7–25),
- the final condition check (line 26) and an eventual choice of the prototype (line 27).

An exhaustive list of potential adjustments and (both interpretation-related and computational) modifications to the strategy introduced above can be found in [64].

We can give the following extended example of how the strategy works:

Example 4. *Let the agent be equipped with a basic mental space* $O = V_{a_1} \times V_{a_2} \times V_{a_3} \times V_{a_4} = \{o_1, \ldots, o_{24}\}$, *where domains of assumed attributes are given as* $V_{a_1} = V_{a_2} = V_{a_3} = \{0,1\}$ *and* $V_{a_4} = \{0,1,2\}$ *(see Definition4). Thus, the basic mental space cognitively available to the agent is a set of vectors of the fixed length* $n = 4$, *e.g.,* $o_1 = [0000]$, $o_2 = [0001]$, $o_3 = [0002]$, $o_4 = [0010]$, $o_5 = [0011]$, $o_6 = [0012]$, $o_7 = [0100]$, $o_8 = [0101]$, $o_9 = [0102]$, $o_{10} = [0110]$, $o_{11} = [0111]$, $o_{12} = [0112]$, $o_{13} = [1000]$, $o_{14} = [1001]$, $o_{15} = [1002]$, $o_{16} = [1010]$, $o_{17} = [1011]$, $o_{18} = [1012]$, $o_{19} = [1100]$, $o_{20} = [1101]$, $o_{21} = [1102]$, $o_{22} = [1110]$, $o_{23} = [1111]$, $o_{24} = [1112]$.

Let us also assume that the basic mental space is enriched with the classic Hamming distance $f_H : O \times O \to \mathbb{N}$, *such that* $\forall o_i, o_j \in O \;\; f_H(o_i, o_j) = \sum_{a \in A} f_{\neq}(o_i[a], o_j[a])$,

where: $f_{\neq}(v,w) = \begin{cases} 0 \;\Leftrightarrow\; v = w \\ 1 \;\Leftrightarrow\; v \neq w. \end{cases}$

Let the Hamming distance f_H *be the chosen function* f *in this example use of Algorithm 1, i.e.,* $f = f_H$.

Finally, let the following set of learning examples be given as $Exp_c = \{1(o_1, is\text{-}c), 1(o_5, is\text{-}c), 1(o_6, not\text{-}c), 2(o_7, is\text{-}c), 2(o_{15}, not\text{-}c), 2(o_{16}, is\text{-}c), 1(o_{17}, not\text{-}c), 1(o_{18}, not\text{-}c)\}$.

Algorithm 1: Prototype-based strategy of learning categories.

Input: *cognitive model* m_c *of the category* c,
set of episodes $Episodes(t)$.
Output: *updated cognitive model* m_c.

$Exp_c := Preprocess(Episodes(t))$;
initialize a chosen prototype as $o_c^\star := NULL$;
$E^+ := E_c^+(Exp_c)$;
$E^- := E_c^-(Exp_c)$;
compute the set $Candidates := extractCandidates(\hat{E}_c^+(Exp_c))$;
while $Candidates \neq \varnothing \wedge o_c^\star = NULL$ **do**
  choose a prototype candidate $o \in Candidates$;
  $Candidates := Candidates \setminus \{o\}$;
  compute distance values $f(o^+, o)$ *for* $o^+ \in E^+$;
  compute distance values $f(o^-, o)$ *for* $o^- \in E^-$;
  $f_{\min}^-(o) := \min_{o^- \in E^-} \{f(o^-, o)\}$;
  $f_{\max}^+(o) := \max_{o^+ \in E^+} \{f(o^+, o)\}$;
  $F^+ := \{f(o^+, o) : o^+ \in E^+ \wedge f(o^+, o) < f_{\min}^-(o)\}$;
  compute a radius of the core $\tau_c^+ := \begin{cases} \max\{f \in F^+\} & F^+ \neq \varnothing \\ NULL & F^+ = \varnothing \end{cases}$;
  $F^- := \{f(o^-, o) : o^- \in E^- \wedge f(o^-, o) > f_{\max}^+(o)\}$;
  compute a radius of the boundary $\tau_c^- := \begin{cases} \min\{f \in F^-\} & F^- \neq \varnothing \\ NULL & F^- = \varnothing \end{cases}$;
  `// Compute a core of potential` c.
  if $\tau_c^+ \neq NULL$ **then**
    $Core_c(o) := \{o^+ : o^+ \in E^+ \wedge f(o^+, o) \leq \tau_c^+\}$
  else
    $Core_c(o) := \varnothing$;
  `// Compute an outer of potential` c.
  if $\tau_c^- \neq NULL$ **then**
    $Outer_c(o) := \{o^- : o^- \in E^- \wedge f(o^-, o) \geq \tau_c^-\}$
  else
    $Outer_c(o) := \varnothing$;
  `// Compute a boundary of potential` c.
  $Boundary_c(o) := (E^+ \cup E^-) \setminus (Core_c(o) \cup Outer_c(o))$;
  if $|Core_c(o)| \geq |Boundary_c(o) \cap E^+|$ **then**
    assign $o_c^\star := o$;
    add a category c with a prototype $o_c^\star$ *and* τ_c^+, τ_c^- *to the ontological knowledge base of the agent;*
if $o_c^\star = NULL$ **then**
  the model m_c *is ill-defined and has not been learned;*

Execution of the strategy for the given input and the given assumptions leads to the following results:

Initial computations

According to definitions from previous sections:

- $E^+ = E_c^+(Exp_c) = \{o_1, o_5, o_7, o_{16}\}$,
- $E^- = E_c^-(Exp_c) = \{o_6, o_{15}, o_{17}, o_{18}\}$,

- $\hat{E}_c^+(Exp_c) = \{1o_1, 1o_5, 2o_7, 2o_{16}\}$.

Let us assume that the centroids of $\hat{E}_c^+(Exp_c)$ are considered as the candidates for prototype. Thus, they need to satisfy the condition given by Equation (14). The set of such objects is as follows:

- $Candidates = \{o_1, o_4\}$.

Iteration 1

Let $o_4 \in Candidates$ be chosen as a candidate for the prototype, i.e., $o = o_4$ and $Candidates = \{o_1\}$. Next, Hamming distance values $f_H(o_4, o^+)$ for all $o^+ \in E^+$ and $f_H(o_4, o^-)$ for all $o^- \in E^-$ are computed, i.e.,

- $f_H(o_4, o_1) = 1, f_H(o_4, o_5) = 1, f_H(o_4, o_7) = 2, f_H(o_4, o_{16}) = 1$,
- $f_H(o_4, o_6) = 1, f_H(o_4, o_{15}) = 3, f_H(o_4, o_{17}) = 2, f_H(o_4, o_{18}) = 2$.

On this basis, we determine in turn $f^-_{\min}(o_4) = 1$, $f^+_{\max}(o_4) = 2$, $F^+ = \varnothing$, $\tau_c^+ = NULL$, $F^- = \{3\}$, and $\tau_c^- = 3$, which leads to the following:

- $Core_c(o_4) = \varnothing$,
- $Outer_c(o_4) = \{o_{15}\}$,
- $Boundary_c(o_4) = \{o_1, o_5, o_6, o_7, o_{16}, o_{17}, o_{18}\}$.

The results achieved do not meet the condition for category c to be accepted because $|Core_c(o_4)| = 0 < 4 = |Boundary_c(o_4) \cap E^+|$. Since the set Candidates is not empty, the next iteration is possible.

Iteration 2

Let the only object $o_1 \in Candidates$ be chosen as a candidate for the prototype, i.e., $o = o_1$ and $Candidates = \varnothing$. Again, Hamming distance values $f_H(o_1, o^+)$ for all $o^+ \in E^+$ and $f_H(o_1, o^-)$ for all $o^- \in E^-$ are computed, i.e.,

- $f_H(o_1, o_1) = 0, f_H(o_1, o_5) = 2, f_H(o_1, o_7) = 1, f_H(o_1, o_{16}) = 2$,
- $f_H(o_1, o_6) = 2, f_H(o_1, o_{15}) = 2, f_H(o_1, o_{17}) = 3, f_H(o_1, o_{18}) = 3$.

On this basis, we determine in turn $f^-_{\min}(o_1) = 2$, $f^+_{\max}(o_1) = 2$, $F^+ = \{0, 1\}$, $\tau_c^+ = 1$, $F^- = \{3\}$, and $\tau_c^- = 3$, which leads to the following:

- $Core_c(o_1) = \{o_1, o_7\}$,
- $Outer_c(o_1) = \{o_{17}, o_{18}\}$,
- $Boundary_c(o_1) = \{o_5, o_6, o_{15}, o_{16}\}$.

In this case, the results achieved meet the condition for category c to be accepted because $|Core_c(o_1)| = 2 \geq 2 = |Boundary_c(o_1) \cap E^+|$. In consequence, the object o_4 is assigned as the prototype $o_c^\star$ to the properly established category c with τ_c^+ and τ_c^- as the category's thresholds. The category c can be integrated with the ontological knowledge base.

5.4. Scheme for Computational Complexity Evaluation

An important question arises about the possible complexity of the actual implementation of the presented strategy. Theorem 1 formulated below concerns this issue:

Theorem 1. *The computational complexity of the Algorithm 1 is of the order:*

$$O(\sum_t |X_t| + |Exp_c| + e + C \cdot p \cdot (|E^+| + |E^-|)) \tag{15}$$

where

- X_t*—a set of objects in* $Episode(t)$,
- e*—computational complexity of the extractCandidates function,*
- $C = |Candidates|$,
- p*—computational complexity of the expression* $f(o_i, o_j)$.

Proof. The first step of the algorithm is preprocessing the data collected by the agent in subsequent episodes to the Exp_c multiset (line 1). The computational complexity of this

step is proportional to the number of objects observed in all the episodes considered. If we denote the set of objects in the $Episode(t)$ by X_t, the computational complexity will be of the order $O(\sum_t |X_t|)$. The next steps of the algorithm (lines 3 and 4) require inspection of the Exp_c multiset in order to find the sets E^+ and E^-. The size of the Exp_c multiset is $|Exp_c| = 2|O|$. In turn, the size of the set O grows exponentially with the size of the set of attributes, because $|O| = \prod_{a \in A} |V_a|$. In a pessimistic case, finding the sets E^+ and E^- can therefore have computational and memory complexity exponentially dependent on $|A|$. In practice, however, we expect that a very small fraction of the elements belonging to Exp_c have the multiplicity $k_o > 0$. Hence, it is convenient to store in the memory only elements of Exp_c for which the multiplicity is $k_o > 0$. This should significantly reduce the memory requirements and the number of operations needed to review the Exp_c multiset.

On line 5, the *extractCandidates* function is executed. Its complexity can be very different depending on the adopted macrostructure and field of application. Here, it is simply denoted by e and treated as a parameter of the formula for the computational complexity of the whole algorithm.

The number of iterations of the **while** loop (line 6) depends on the number of candidates. We will abbreviate it with $C = |Candidates|$. In the worst case, it will be equal to the size of the O set, but in practice the *extractCandidates* function should return a much smaller set of candidates.

Inside the loop, in line 9 the macrostructure value $f(o, o^+)$ needs to be calculated for all $o^+ \in E^+$. Again, the calculation of $f(o, o^+)$ strongly depends on the adopted macrostructure and the field of application. Denoting by p the computational complexity of the expression $f(o_i, o_j)$, to complete the instruction from line 9, $p \cdot |E^+|$ operations need to be executed. Similarly, it takes $p \cdot |E^-|$ operations to execute instructions from line 10.

The complexity of the other instructions inside the loop (lines 11 to 28) is linearly dependent on the size of the sets E^+ and E^-. In summary, the computational complexity of the **while** loop is of the order $O(C \cdot p \cdot (|E^+| + |E^-|))$.

The final complexity of the entire algorithm results from the summation of the above estimates. □

The Theorem 1 shows that the computational complexity of the Algorithm 1 can be polynomial with respect to the number of objects observed by the agent in all episodes $\sum_t |X_t|$ and the number of attributes describing them $|A|$, provided that the following conditions are met:

- the size of the *Candidates* set does not grow exponentially with the number of attributes,
- the computational complexity of the expression $f(o_i, o_j)$ does not increase exponentially with the number of attributes,
- the computational complexity of the *extractCandidates* function does not increase exponentially with the number of objects or attributes,
- the size of the practically used part of the Exp_c multiset (i.e., elements with the multiplicity $k_o > 0$) does not increase exponentially with the number of attributes.

If any of the above conditions is not met, then the complexity of the strategy becomes exponential and therefore might be relatively hard to be effectively applied in practice. The complexity of *extractCandidates* seems to be the key parameter, in particular because, at the design level, in many implementations it may involve the need to solve the choice problem in a mathematical space. As has long been proven, the latter problem usually belongs to the class of tasks with exponential computational complexity, depending mainly on adopted choice criteria, but at a deeper level on the structure of objects and related metrics (distance/similarity function f). A sketch of a general theoretical structure from which detailed implementation models of *extractCandidates* can be derived, and then within

which the analysis of the complexity of a selected implementation model can be carried out, can be found, for example, in chapter 3 of [65]: *Consensus as a Tool for Conflict Solving*.

6. Cognitive Semantics

The above presentation of the details and description of the pragmatic function of individual elements of the agent's architecture allows us to move onto an explicit definition of the cognitive semantics of the considered cases of autoepistemic membership statements. Formulating the definition of cognitive semantics will involve establishing a model of reference for each of these cases, assuming that reference is understood as in the semiotic triangle (see Figure 1). The main conceptual elements involved in defining the cognitive semantics for an individual autoepistemic statement are shown in Figure 6, which refers to the representation of the architecture given in Figure 2.

The presented statement $Know(x_{n,5} \in c)$ is used to communicate the agent's belief regarding the membership of an external object marked in Figure 6 as referent. The object was observed in relation to time point t_n. Therefore the internal model of the object is a component of $Episode(t_n)$. To emphasize the assumption that this episode is interpreted as a representation of an actual (current) observed state of the external world, it is still present in the working memory.

Graphically, an edge represents the relationship of object $x_{n,5}$ and its internal mental reflection within the agent, represented as element o_{20} in the agent's basic mental space. It creates a possibility for the agent to internally analyze a correlation between $x_{n,5}$ and the model of a particular category c. The resulting location of $x_{n,5}$ within the scope of the model for the category may vary: in the core, the border, or the outer sphere of the category.

The occurrence of $x_{n,5}$ in an episode, the assignment of this episode to the role of a model of a currently existing state of the external world, the location of object $x_{n,5}$ in the models of categories stored in the assumed repository, and some additional numeric characteristics of this location (to be presented further) are the key components of the way in which references are defined while being part of cognitive semantics.

The overall state of the above-mentioned elements, determined to the greatest extent by the adapted category models and the episode model, will be called the state of cognition and for formal purposes determined as follows:

Definition 7. *At each time point $t \in T$, the t-related state of cognition of the agent is described by the following tuple*

$$SP(t) = (M(t), Episode(t)) \tag{16}$$

where $M(t)$ is a set of models of categories stored in the agent's ontology and $Episode(t)$ is an episode at time point t available to the agent's perception.

In turn, the symbol $\models_G$ used in the figure represents the so-called epistemic satisfaction relation. The conditions for this relationship to be held will also be the definition of cognitive semantics of statement $Know(x_{n,5} \in c)$. A similar approach and theoretical concepts were used to define the cognitive semantics of the other autoepistemic statements considered in previous stages of our related research (e.g., [18]).

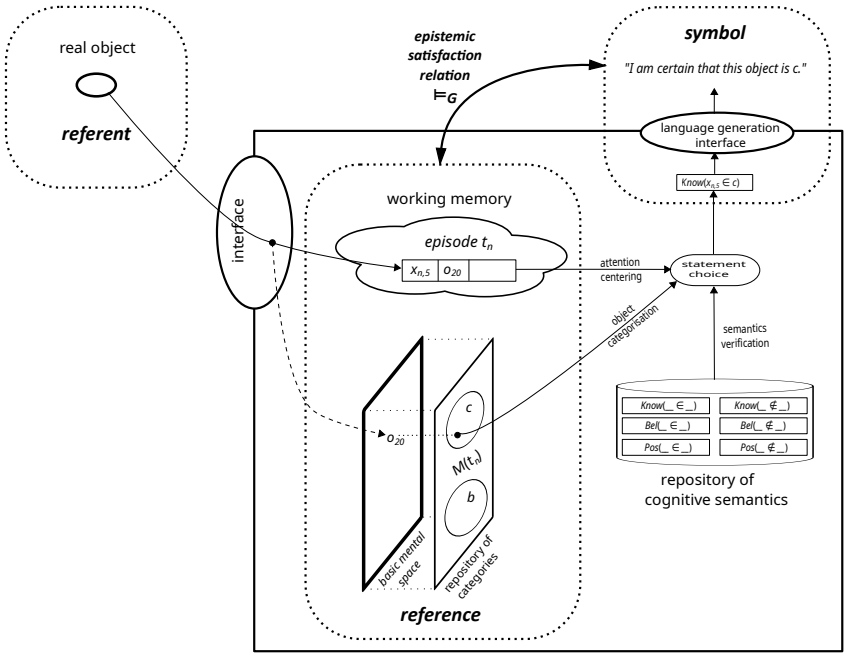

Figure 6. Epistemic satisfaction relation.

In the extension of cognitive semantics of autoepistemic membership statements (let us recall that the latter originally was considered only for the case of categories without prototypes), the state of cognition, in which the extended statements considered in this paper are anchored, includes an additional dimension of knowledge representation, namely the phenomenon of blurring the boundaries of categories. Therefore, in general, when the agent designates its attitude towards membership of an object x in a category c, the first step is to determine in which region of the mental model of category c the object x is located. If it is in the core or in the outer region (see Section 4.3), the case seems to be rather obvious and simple from the commonsense point of view. If it is in the boundary, which is the region representing socially originated uncertainty in the category's structure (just mentioned by "blurry boundary"), then additional conditions must be checked. This general assumption is further taken into account in definitions of particular cognitive semantics.

We consider the core of the category's model m_c to include objects that most certainly belong to the category c. Therefore, including an object in the core of the category is the basis for grounding the statement, the intuitive meaning of which is represented by $Know(x \in c)$. This intuition is captured in the following simple definition:

Definition 8. *Let the time point t and the state of cognitive processes $SP(t)$ described by the episode $Episode(t)$ and the set of cognitive models $M(t)$ containing the well-defined model m_c be given. For each object $x \in X_t$ and category c, we assume that the epistemic satisfaction relation $SP(t) \models_G Know(x \in c)$ holds if and only if*

$$f(o, o_c^\star) \leq \tau_c^+ \tag{17}$$

where object x is observed realization of mental object o at time point t and $o_c^\star$ is a prototype of c.

We recognize that the outer region of the category's model m_c includes objects that are definitely excluded from the category c. Therefore, including an object in the outer region of the category is the basis for grounding the statement, the intuitive meaning of which can be represented by $Know(x \notin c)$. The related definition is as follows:

Definition 9. *Let the time point t and the state of cognitive processes $SP(t)$ described by the episode $Episode(t)$ and the set of cognitive models $M(t)$ containing the well-defined model m_c be*

given. For each object $x \in X_t$ and category c, we assume that the epistemic satisfaction relation $SP(t) \models_G Know(x \notin c)$ holds if and only if

$$f(o, o_c^\star) \geq \tau_c^- \tag{18}$$

where object x is observed realization of mental object o at time point t and $o_c^\star$ is a prototype of c.

We consider that the boundary of the category's model m_c includes objects that may or may not belong to the category c (due to the fact that the social attitude to such membership varied and was not conclusive within the population). Therefore, including an object in the boundary of categories is the basis for establishing modal statements with operators of belief and possibility, the intuitive meaning of which can be represented by $Bel(x \in c)$ and $Pos(x \in c)$, respectively.

Unlike the case of the *Know* operator, which is based purely on a distance from the prototype, the current situation where the object is located within the boundary of c gives the agent a more vague feeling which demands a more intense analysis. In consequence, the agent focuses on particular pieces of learning experience which positively correlate with the current situation (based on a distance from a mental reflection o of the object x). It focuses on an area of the mental space surrounding the currently analyzed object and tries to evaluate an overall impact of the learning experience related to the observed object and category c.

To formally define this impact, similarly to the approach developed and verified in [18], we will use the concept of so-called relative grounding strength, determined by the distance of the considered object from both positive and negative pieces of previously collected learning examples located in the basic mental space in the immediate surroundings of this object. Such surroundings are called an epistemic neighborhood:

Definition 10. *For a given object $o \in O$, by an epistemic neighborhood EN_c we understand a set of objects defined as follows:*

$$EN_c(o, \varepsilon) = \{e \in (E^+(Exp_c) \cup E^-(Exp_c)) : f(e, o) \leq \varepsilon\} \tag{19}$$

where $\varepsilon \in \mathbb{R}_{\geq 0}$ is called the radius of the epistemic neighborhood.

The radius of the epistemic neighborhood ε can be determined in various ways; for example, it can be an experimentally chosen constant. In this paper, we propose that it is determined by the function ER, depending on the value of thresholds τ_c^- and τ_c^+ delineating the category regions. In the following considerations, we will assume that the value of the ER function depends linearly on the width of the boundary of the category model, i.e., $ER(\tau_c^-, \tau_c^+) = \alpha(\tau_c^- - \tau_c^+)$, where $\alpha \in \mathbb{R}_+$ is an assumed positive coefficient of the radius of the epistemic neighborhood. Following this formula, the greater the boundary of the category, and hence the greater the uncertainty as to whether an observed object belongs to a category, the greater the epistemic neighborhood considered when grounding statements. A larger neighborhood usually means that the decision to select a modal operator will be made on the basis of more experience.

The following definition introduces the concept of the relative grounding strength tailored to our extension:

Definition 11. *For a set of objects $Q \subseteq (E^+(Exp_c) \cup E^-(Exp_c))$ the relative grounding strength $\lambda_c(Q)$ is defined as follows:*

$$\lambda_c(Q) = \begin{cases} 0 & if \quad |Q| = 0 \\ \frac{|Q \cap E^+(Exp_c)|}{|Q|} & if \quad |Q| > 0. \end{cases} \tag{20}$$

As was thoroughly (analytically) proven in [18], a value of the relative grounding strength, along with the so-called modality thresholds λ_{minPos}, λ_{maxPos}, λ_{minBel} and λ_{maxBel}, constitutes an effective numerical tool for determining the ranges of relative grounding strength values associated with the proper (commonsense coherent) use of operators *Bel* and *Pos*. Example values of such thresholds are explicitly given in [18]. In this study, we adopt these concepts, but we also include (for simplicity of presentation) that it will be sufficient to concentrate only on λ_{minBel}. Such deliberate simplification is adopted in the following definitions:

Definition 12. *Let the time point t, the state of cognitive processes $SP(t)$ described by the episode $Episode(t)$, the set of cognitive models $M(t)$ containing the well-defined model m_c, the radius of the epistemic neighborhood ε, and the $\lambda_{minBel} \in (0,1]$ threshold be given. For any object $x \in X_t$ and category c, we assume that epistemic satisfaction relations $SP(t) \models_G Bel(x \in c)$ and $SP(t) \models_G Pos(x \notin c)$ hold if and only if*

$$\left(\tau_c^+ < f(o, o_c^\star) < \tau_c^-\right) \wedge \left(\lambda_c(EN_c(o,\varepsilon)) \geq \lambda_{minBel}\right) \tag{21}$$

where object x is observed realization of mental object o at time point t and $o_c^\star$ is a prototype of c.

Definition 13. *Let the time point t, the state of cognitive processes $SP(t)$ described by the episode $Episode(t)$, the set of cognitive models $M(t)$ containing the well-defined model m_c, the radius of the epistemic neighborhood ε, and the $\lambda_{minBel} \in (0,1]$ threshold be given. For any object $x \in X_t$ and category c, we assume that epistemic satisfaction relations $SP(t) \models_G Bel(x \notin c)$ and $SP(t) \models_G Pos(x \in c)$ hold if and only if*

$$\left(\tau_c^+ < f(o, o_c^\star) < \tau_c^-\right) \wedge \left(\lambda_c(EN_c(o,\varepsilon)) < \lambda_{minBel}\right) \tag{22}$$

where object x is observed realization of a mental object o at time point t and $o_c^\star$ is a prototype of c.

The following examples illustrate the way in which the above definitions of cognitive semantics can shape language production:

Example 5. *Let us suppose that for the model of category c, the thresholds are equal to $\tau_c^+ = 5$ and $\tau_c^- = 8$. In $Episode(t_3)$, two objects $x_{3,8}$ and $x_{3,9}$ appeared in the agent's range of perception. Suppose that in the agent's cognition process the object $x_{3,8}$ from working memory corresponds to the object o_8 in embodied ontology, and the object $x_{3,9}$ from working memory corresponds to the object o_9 in embodied ontology. The distances between the objects and the prototype are $f(o_8, o_c^\star) = 3$ and $f(o_9, o_c^\star) = 10$. The above situation is presented in Figure 7.*

Since $f(o_8, o_c^\star) = 3 \leq \tau_c^+ = 5$ according to Definition 8, an epistemic satisfaction relation holds for formula $Know(x_8 \in c)$ and such a formula could be generated by the agent. The intuitive meaning of the formula can be expressed as "I know that object $x_{3,8}$ belongs to category c".

Since $f(o_9, o_c^\star) = 10 \geq \tau_c^- = 8$ according to Definition 9, an epistemic satisfaction relation holds for formula $Know(x_9 \notin c)$ and such a formula could be generated by an agent. The intuitive meaning of the formula can be expressed as "I know that object $x_{3,9}$ does not belong to category c".

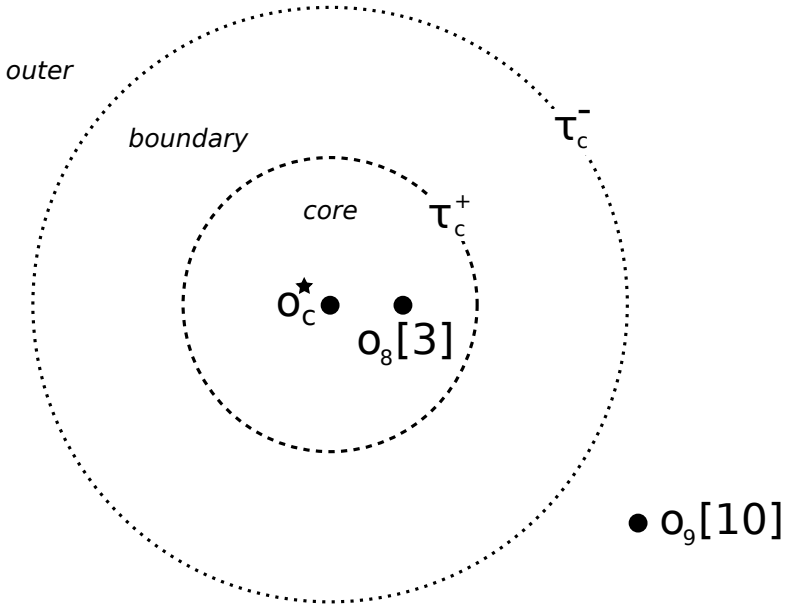

Figure 7. Grounding of objects in core and outer region of category model.

Example 6. *A more complicated case is when an object is included in the boundary of the category. In order to establish the right statement, the agent must then compare the considered object not only with the prototype but also with other objects in embodied ontology. Let us make similar assumptions as in the previous example, except that this time the distances between the objects and the prototype are* $f(o_8, o_c^\star) = 6$ *and* $f(o_9, o_c^\star) = 6$*. The above situation is presented in Figure 8.*

Since $\tau_c^+ = 5 < f(o_8, o_c^\star) = 6 < \tau_c^- = 8$ *and* $\tau_c^+ = 5 < f(o_9, o_c^\star) = 6 < \tau_c^- = 8$*, the relative grounding strength must be determined for both objects. The first step is to establish the radius of the epistemic neighborhood* ε*. As already mentioned, we apply the formula for the linear dependence of the radius on the width of the boundary. Assuming* $\alpha = 0.8$*, we obtain* $\varepsilon = ER(\tau_c^-, \tau_c^+) = \alpha(\tau_c^- - \tau_c^+) = 0.8 \cdot (8 - 5) = 2.4$

According to Definition 10, we determine the epistemic neighborhood of objects, i.e., $EN_c(o_8, \varepsilon)$ *and* $EN_c(o_9, \varepsilon)$*. For this purpose, it is necessary to calculate the distance from the objects* o_8 *and* o_9 *to the objects in embodied ontology in a model of c. Let us assume that the above distances are given as in Table 6 and that furthermore:*

- $E^+(Exp_c) = \{o_1, o_2, o_3, o_5\}$
- $E^-(Exp_c) = \{o_4, o_6, o_7\}$.

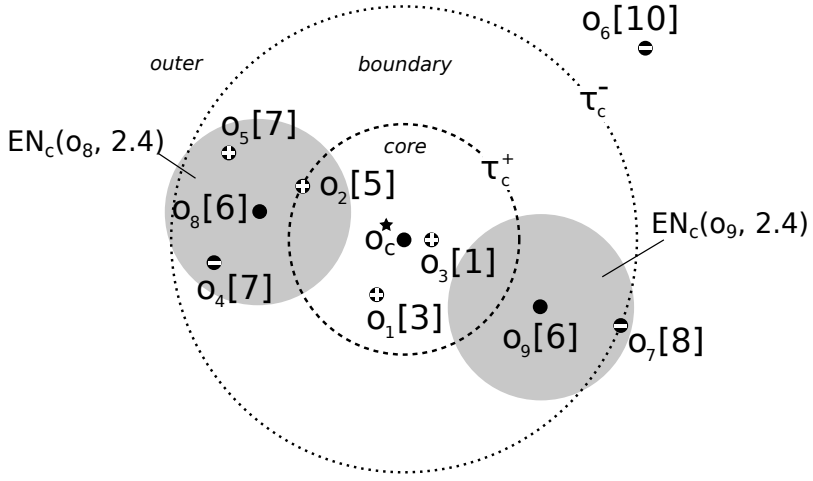

Figure 8. Grounding of objects in boundary of category model.

Table 6. Distances between objects in example.

o_i	o_1	o_2	o_3	o_4	o_5	o_6	o_7
$f(o_8, o_i)$	6	1	7	1	1	14	14
$f(o_9, o_i)$	6	11	5	11	13	8	2

Based on the above data, we determine $EN_c(o_8, \varepsilon) = EN_c(o_8, 2.4) = \{o_2, o_4, o_5\}$ *and* $EN_c(o_9, \varepsilon) = EN_c(o_9, 2.4) = \{o_7\}$. *Following Definition 11, we can calculate the relative grounding strength* $\lambda_c(EN_c(o_8, 2.4)) = \frac{|\{o_2, o_5\}|}{|\{o_2, o_4, o_5\}|} = 2/3$, *and also we can calculate that* $\lambda_c(EN_c(o_9, 2.4)) = \frac{|\varnothing|}{|\{o_7\}|} = 0/1$.

Let us assume $\lambda_{minBel} = 0.5$. *Such a threshold value means that if at least half of the elements in the epistemic neighborhood of the considered object are positive experiences, then the agent will be willing to establish a statement with the operator of the belief that the object belongs to the category. On the other hand, if in the epistemic neighborhood of the considered object more than half of the elements are negative experiences, then the agent will be willing to establish a statement with the operator about the possibility regarding the belonging of the object to the category.*

Since $\lambda_c(EN_c(o_8, 2.4)) = 2/3 \geq \lambda_{minBel} = 0.5$, *then according to Definition 12, the epistemic satisfaction relation holds for formulas* $Bel(x_8 \in c)$ *and* $Pos(x_8 \notin c)$, *and such formulas could be generated by an agent. The intuitive meaning of the formulas can be expressed as "I believe that object* $x_{3,8}$ *belongs to category c." and "It is possible that object* $x_{3,8}$ *does not belong to category c".*

Since $\lambda_c(EN_c(o_9, 2.4)) = 0 < \lambda_{minBel} = 0.5$ *according to Definition 13, an epistemic satisfaction relation holds for formulas* $Bel(x_9 \notin c)$ *and* $Pos(x_9 \in c)$, *and such formulas could be generated by an agent. The intuitive meaning of the formulas can be expressed as "I believe that object* $x_{3,9}$ *does not belong to category c." and "It is possible that object* $x_{3,9}$ *belongs to category c".*

Note that the distance from the o_8 *and* o_9 *objects to the category's prototype is the same, but the statements generated by the agent are different due to the different neighborhood of each object.*

7. Verification of the Model

Probably one of the most interesting features of our proposed method of introducing computational mechanisms for the production of autoepistemic membership statements is the resulting possibility of carrying out analytical verification of the features of this process, when already at the stage of determining the agent's architecture and adopting specific definitions of cognitive semantics. Indeed, this possibility was already used to handle autoepistemic membership statements communicating beliefs about the belonging of objects to categories without a prototype [18–21,26,27]. Below we present, in our opinion, the most important features of our proposed extended cognitive semantics mechanism, emphasizing at the same time that the list of theorems given below should be considered together with the theorems formulated, among others, in work [18]. In this way, a broader description of the artificial agent's linguistic behavior is obtained, along with a justification that this behavior meets the commonsense constraints placed on the behavior in natural language discourse.

To simply illustrate the pragmatics of such commonsense constraints, we can give the following cases:

- it should not be allowed to utter certain autoepistemic membership statements simultaneously about the same object, e.g., it is not acceptable for the agent to generate the following statements in relation to the same episode, as they would be considered nonsensical/contradictory by other participants of communication:
  - "I know that object x belongs to category c",
  - "I know/I believe/It is possible that object x does not belong to category c",
- it should be allowed to utter certain statements simultaneously regarding the same object, e.g., it is permissible for the agent to generate the following statements in relation to the same episode:

– "I believe that object x belongs to category c",
– "It is possible that object x does not belong to category c".

Such properties are captured by theorems presented below, along with the proofs based on the definitions of epistemic satisfaction relations.

Theorems 2 and 3 concern a fairly obvious commonsense limitation, that an agent should not make statements indicating that it knows that an object both belongs to and does not belong to category c.

Theorem 2. *If relation* $SP(t) \models_G Know(x \in c)$ *holds, then relation* $SP(t) \models_G Know(x \notin c)$ *does not hold.*

Proof. The epistemic satisfaction relation $SP(t) \models_G Know(x \in c)$ holds (Definition 8) if and only if

$$f(o, o_c^\star) \leq \tau_c^+ \tag{23}$$

where x is the observed realization of mental object o at time point t. In previous sections, we assumed that for a well-defined model it is always $\tau_c^+ < \tau_c^-$ (condition (7)). It follows that

$$f(o, o_c^\star) < \tau_c^-. \tag{24}$$

Thus, the condition $f(o, o_c^\star) \geq \tau_c^-$ required for the epistemic satisfaction relation $SP(t) \models_G Know(x \notin c)$ is not fulfilled (Definition 9). □

Theorem 3. *If relation* $SP(t) \models_G Know(x \notin c)$ *holds, then relation* $SP(t) \models_G Know(x \in c)$ *does not hold.*

Proof. The epistemic satisfaction relation $SP(t) \models_G Know(x \notin c)$ holds (Definition 9) if and only if

$$f(o, o_c^\star) \geq \tau_c^- \tag{25}$$

where x is the observed realization of mental object o at time point t. In previous sections, we assumed that for a well-defined model it is always $\tau_c^+ < \tau_c^-$ (condition (7)). It follows that

$$f(o, o_c^\star) > \tau_c^+. \tag{26}$$

Thus, the condition $f(o, o_c^\star) \leq \tau_c^+$ required for the epistemic satisfaction relation $SP(t) \models_G Know(x \in c)$ is not fulfilled (Definition 8). □

The next group containing Theorems (4–7) deals with situations where an agent generates statements indicating that it knows that an object belongs to category c or that it does not belong to category c. In both cases, the agent should not produce simultaneously statements with weaker confidence about the class membership of an object.

Theorem 4. *If relation* $SP(t) \models_G Know(x \in c)$ *holds, then*

- *relation* $SP(t) \models_G Bel(x \in c)$ *does not hold,*
- *relation* $SP(t) \models_G Bel(x \notin c)$ *does not hold,*
- *relation* $SP(t) \models_G Pos(x \in c)$ *does not hold,*
- *relation* $SP(t) \models_G Pos(x \notin c)$ *does not hold.*

Proof. The epistemic satisfaction relation $SP(t) \models_G Know(x \in c)$ holds (Definition 8) if and only if

$$f(o, o_c^\star) \leq \tau_c^+ \tag{27}$$

where x is the observed realization of mental object o at time point t.

Thus, the condition $\tau_c^+ < f(o, o_c^\star)$ required for epistemic satisfaction relations $SP(t) \models_G Bel(x \in c)$ and $SP(t) \models_G Pos(x \notin c)$ is not fulfilled (Definition 12). The same condition is

required for epistemic satisfaction relations $SP(t) \models_G Bel(x \notin c)$ and $SP(t) \models_G Pos(x \in c)$ (Definition 13). □

Theorem 5. *If any of the following relationships hold*

- $SP(t) \models_G Bel(x \in c)$,
- $SP(t) \models_G Bel(x \notin c)$,
- $SP(t) \models_G Pos(x \in c)$,
- $SP(t) \models_G Pos(x \notin c)$,

then relation $SP(t) \models_G Know(x \in c)$ *does not hold.*

Proof. This theorem is the contraposition of Theorem 4. □

Theorem 6. *If relation* $SP(t) \models_G Know(x \notin c)$ *holds, then*

- *relation* $SP(t) \models_G Bel(x \in c)$ *does not hold,*
- *relation* $SP(t) \models_G Bel(x \notin c)$ *does not hold,*
- *relation* $SP(t) \models_G Pos(x \in c)$ *does not hold,*
- *relation* $SP(t) \models_G Pos(x \notin c)$ *does not hold.*

Proof. The epistemic satisfaction relation $SP(t) \models_G Know(x \notin c)$ holds (Definition 9) if and only if

$$f(o, o_c^\star) \geq \tau_c^- \tag{28}$$

where x is the observed realization of mental object o at time point t.

Thus, the condition $f(o, o_c^\star) < \tau_c^-$ required for epistemic satisfaction relations $SP(t) \models_G Bel(x \in c)$ and $SP(t) \models_G Pos(x \notin c)$ is not fulfilled (Definition 12). The same condition is required for epistemic satisfaction relations $SP(t) \models_G Bel(x \notin c)$ and $SP(t) \models_G Pos(x \in c)$ (Definition 13). □

Theorem 7. *If any of the following relationships hold*

- $SP(t) \models_G Bel(x \in c)$,
- $SP(t) \models_G Bel(x \notin c)$,
- $SP(t) \models_G Pos(x \in c)$,
- $SP(t) \models_G Pos(x \notin c)$,

then relation $SP(t) \models_G Know(x \notin c)$ *does not hold.*

Proof. This theorem is the contraposition of Theorem 6. □

If the agent generates a statement indicating that it believes that an object belongs (or does not belong) to category c, then it should not produce at the same time a statement with weaker confidence. Nor should it produce a statement indicating that it believes an opposite state of membership. The correct behavior of an agent in the above situations is proven for Theorems 8 and 9.

Theorem 8. *If relation* $SP(t) \models_G Bel(x \in c)$ *holds, then*

- *relation* $SP(t) \models_G Bel(x \notin c)$ *does not hold,*
- *relation* $SP(t) \models_G Pos(x \in c)$ *does not hold.*

Proof. The epistemic satisfaction relation $SP(t) \models_G Bel(x \in c)$ holds (Definition 12) if and only if

$$\left(\tau_c^+ < f(o, o_c^\star) < \tau_c^-\right) \wedge \left(\lambda_c(EN_c(o, \varepsilon)) \geq \lambda_{minBel}\right) \tag{29}$$

where x is the observed realization of mental object o at time point t and $EN_c(o, \varepsilon)$ is the epistemic neighborhood of the object o with radius ε.

Thus, the condition $\lambda_c(EN_c(o, \varepsilon)) < \lambda_{minBel}$ required for epistemic satisfaction relations $SP(t) \models_G Bel(x \notin c)$ and $SP(t) \models_G Pos(x \in c)$ is not fulfilled (Definition 13). □

Theorem 9. *If relation* $SP(t) \models_G Bel(x \notin c)$ *holds, then*

- *relation* $SP(t) \models_G Bel(x \in c)$ *does not hold,*
- *relation* $SP(t) \models_G Pos(x \notin c)$ *does not hold.*

Proof. The epistemic satisfaction relation $SP(t) \models_G Bel(x \notin c)$ holds (Definition 13) if and only if

$$\left(\tau_c^+ < f(o, o_c^\star) < \tau_c^-\right) \wedge \left(\lambda_c(EN_c(o, \varepsilon)) < \lambda_{minBel}\right) \tag{30}$$

where x is the observed realization of mental object o at time point t and $EN_c(o, \varepsilon)$ is the epistemic neighborhood of the object o with radius ε.

Thus, the condition $\lambda_c(EN_c(o, \varepsilon)) \geq \lambda_{minBel}$ required for epistemic satisfaction relations $SP(t) \models_G Bel(x \in c)$ and $SP(t) \models_G Pos(x \notin c)$ is not fulfilled (Definition 12). □

Analogically to the above, if the agent generates a statement indicating that it is possible that an object belongs (or does not belong) to category c, then it should not produce at the same time a statement with stronger confidence—Theorems 10 and 11.

Theorem 10. *If relation* $SP(t) \models_G Pos(x \in c)$ *holds, then relation* $SP(t) \models_G Bel(x \in c)$ *does not hold.*

Proof. This theorem is the contraposition of the second part of Theorem 8. □

Theorem 11. *If relation* $SP(t) \models_G Pos(x \notin c)$ *holds, then relation* $SP(t) \models_G Bel(x \notin c)$ *does not hold.*

Proof. This theorem is the contraposition of the second part of Theorem 9. □

If the agent expresses the belief that an object belongs to some category c, it is rational that the agent accepts the possibility that this object does not belong to category c. It should therefore be possible for the agent to express both of the above opinions at the same state of knowledge. Such pairs of statements could potentially be connected with an additional language connector (not defined formally in this work), e.g., "I believe that object x belongs to category c, however, it is possible that object x does not belong to category c".

Similarly, if the agent expresses the belief that an object does not belong to some category c, it is rational that the agent accepts the possibility that this object does belong to category c. The next two theorems are proven for the above situations.

Theorem 12. *Relations* $SP(t) \models_G Bel(x \in c)$ *and* $SP(t) \models_G Pos(x \notin c)$ *hold in the same state of knowledge.*

Proof. It follows directly from Definition 12, where the conditions of the epistemic satisfaction relation are the same for both formulas. □

Theorem 13. *Relations* $SP(t) \models_G Bel(x \notin c)$ *and* $SP(t) \models_G Pos(x \in c)$ *hold in the same state of knowledge.*

Proof. It follows directly from Definition 13, where the conditions of the epistemic satisfaction relation are the same for both formulas. □

The general conclusion resulting from the list of theorems formulated above, supplemented by the theorems discussed in work [18], is that the design and implementation of an artificial cognitive agent as is described by the proposed theory (including specific definitions of cognitive semantics) makes it possible to realize an intelligent interactive system generating linguistic behavior consistent with human linguistic behavior, at least in terms of processing autoepistemic membership statements communicating beliefs about the belonging of observed objects to a category with a prototype. After taking into account the results presented in works [19–27] (some of which also presented simulation studies),

the given conclusion can be extended to the case of autoepistemic extensions of more complex membership statements.

8. Results and Discussion

It is naturally desirable to add natural-language-oriented capabilities to technical systems and intelligent interfaces. Increasing use of well-performing black box models for language generation clashes against postulated transparency (including traceability, explainability, and communication) of *Trustworthy AI* [66] required in order to employ such solutions in critical applications like national security or healthcare. As shown in the field of eXplainable Artificial Intelligence (XAI), reverse engineering of interpretable models for existing black box models is not feasible in general and interpretable models should be designed from the ground up.

This article has shown a fully interpretable approach to a computationally realized process of producing autoepistemic membership statements along with the underlying process of category formation (for the case of categories with a prototype). We follow discussed commonsense properties of human behavior and try not to stop at a question of "*How it works?*" but also provide a clear answer to a question of "*Why is it designed like this?*".

The aim of this article was to develop and analyze the model of a system that can imitate human behavior in terms of generating statements about the membership of an observed object to a particular category with a prototype. Harnad's theory of grounding [40] shows that a symbol's meaning cannot be purely external to the agent and this theoretical assumption has been adopted and reproduced in many dimensions of our approach.

Recalling the infeasibility of an overly strict criterial-attribute model [10], our attention has been successfully shifted to Rosch's theses formulated for the prototype semantics [13,14]. On this basis, the cognitive model including categories with a prototype has been defined, and the cognitive agent's architecture has been developed, allowing for the learning of the category model. This prototype semantics has been effectively combined in our work with the concept of conceptual spaces [45–47]. We concluded that, with an application of a distance-based model of a category with a prototype, the following of Rosch's theses presented in Section 1.2 are realizable:

1. The category has an internal prototype structure.
2. The degree of representativeness of a given item needs to correspond to the degree of its membership to a category. In our model, belonging to a category is determined, among other things, on the basis of the distance from the prototype. The more representative elements are those closer to the prototype, and they are more likely to be included in the category, including its core.
3. The elements of a given category do not have to possess properties common to all elements. In our model, category elements are connected to the prototype. The model can be extended in the future to include connections between elements.
4. The boundaries of categories or concepts ought to be fuzzy. In our model, the category boundary contains elements that may or may not belong to a category.
5. The belonging to a given category needs to be based on the degree of similarity to the prototype. In our work, the measures of distance from the prototype are considered, due to their easier implementation. However, distance can simply be thought of as the inverse measure of similarity, and so they can be used interchangeably.
6. The belonging to a category should not be determined only in an analytical manner, but also in a holistic manner. In our model, one does not analyze sets of necessary and sufficient conditions for the attributes of objects, like in classical definitions of categories. Instead, a more holistic measure of distance to the prototype is applied.

The proposed model allows us to execute some additional cognitive phenomena similar to the ones studied by Rosch, e.g., the agent could rank the elements belonging to the category according to their degree of representativeness (distance from the prototype). We do not describe related experiments in detail, due to lack of space, but it is an interesting

property, testifying that the properties of our model are similar to the ones of cognitive processes in humans.

The presented model of cognitive semantics, tailored to categories with a prototype, substantially extends previous results for categories without a prototype [18–27]. It has been proven that the proposed cognitive semantics have properties necessary to conform to human linguistic behavior. This is important for two reasons. Firstly, it shows an example of the practical application of the model of a category with a prototype. Secondly, it shows that the use of a model reflecting a human category processing makes it possible to build subsequent layers and processing modules within the artificial agent, which will also be consistent with human behavior, thus fulfilling commonsense expectations regarding the agent's human-like behavior.

The main goal of this article was to propose and sufficiently describe a substantial extension of the original model presented in the series of papers [18–27] by adding mechanisms for handling categories with prototypes. Therefore, the presentation concentrates only on the proposed extensions of the agent's architecture and the related reformulations of original cognitive semantics. At the same time, any extended presentation of these elements of the proposed mechanism, which are subject to variants in particular implementations, has been deliberately omitted. This omission concerns in particular any detailed presentation of already prepared and verified implementations of the function for prototype candidate selection, which is an important step in the overall strategy for managing prototypes. As is well known, choosing candidates for possible prototypes might be a computationally complex task, depending on the similarity (distance) measures and choice criteria used [65]. Therefore, a proper and satisfactory presentation of the mentioned implementations requires separate and extensive analysis and discussion.

To sum up, the proposed solution is technically feasible. Importantly, the presented model is easily adopted by users thanks to the application of semi-natural language, and it is transparent for designers. Moreover, it even has elements that self-diagnose the quality of the obtained category model, rejecting the models that do not meet the acceptance condition.

Directions of Future Research

It should be obvious to the reader that any artificial computing system is usually an imperfect and often deliberately simplified approximation of its natural prototype. As this is also the case with the architecture proposed above, we would like to outline the following general directions for further development of the presented solution (both in terms of potential additional elements of the very model itself and in terms of the handled types of statements).

Firstly, this article discusses the agent's cognitive competences regarding acquiring and processing beliefs about the states of objects in only one base mental space. This limitation is only of an editorial nature. It should be remembered that designing agents for real practical contexts is almost always associated with the need to equip them with cognitive competences (including linguistic competences) relating to more than one basic mental space. Each of such spaces then corresponds to a different type of atomic object in the real world. Moreover, more complete and externally grounded models of embedded agent ontology should also provide for the possibility of conceptualizing the space of complex objects, i.e., those that can be decomposed into a collection of two or more atomic objects. It is quite obvious that expanding the microstructure of objects in the mental spaces operated by the agent always involves the need to introduce more complex measures of distance or similarity. As a result, the methods of learning categories with a prototype need to be tailored to the particular implementation.

Secondly, this article introduces a category model with a prototype equipped with a spherical core of the category. Adopting such a solution is sufficient to maintain the logical and commonsense coherence of the argumentation presented in this article. It seems interesting to investigate more flexible definitions of the cores and to allow for solutions in which the core takes a form of, for example, a convex figure with a clearly indicated

center acting as a prototype. An analogous postulate can also be formulated regarding the definition of the epistemic neighborhood.

Thirdly, this article focuses on a specific class of models and computational methodologies suitable for realizing specific and effective implementations of the proposed agent. It should be noted that the suggested models and implementation methodologies belong to the group of classical computational tools that can be easily applied to the case of the agent's sub-systems with symbolic and relational representation of knowledge. This is particularly visible in the examples proposed in the text of this article. However, the increasing availability of powerful computers expands the spectrum of implementation tools that can be used, e.g., with deep learning (connectionist) techniques requiring significant, but currently already available, computational power. Developing a way to use these techniques to implement the process of learning for a model of a category with a prototype, and in the longer term also to induce the basic mental spaces themselves, seems to be an interesting and potentially very valuable direction for future research. Regarding effective implementations, it is also important to further investigate a way in which the whole agent uses the category learning procedure. Categories need to be periodically evaluated against new learning samples and in terms of their internal consistency. This results in a potential need for an agent to selectively re-initiate a process of category formation for a subset of previously learned categories. Some insights on the maintenance of the categorization can be drawn from such works as Xu et al. [67].

Fourthly, the model presented in this article assumes complete observational information regarding the state of the observed object, which from the point of view of many applications is an unacceptable idealization. This issue, although it has already been the subject of our considerations, has not been presented here purely due to the editorial constraints.

Fifthly, the possible inclusion in the model of a specific agent of the possibility of representing many mental spaces and defining sets of conceptual categories (including those with a prototype) naturally directs attention to the possibility of asking questions about the semantic relations between the categories cognitively available to a given agent. Examples of such relationships are synonymy, antonymy, subsumption, etc. As one can expect, the appearance of new structural elements at the level of knowledge representation will expand the spectrum of elements taken into account when defining the cognitive semantics of the considered class of autoepistemic statements.

9. Conclusions

This article shows that it is possible to design an interactive cognitive agent which, in terms of the processing of autoepistemic membership statements, will generate linguistic behavior consistent with the commonsense expectations of a natural language user. This main conclusion builds upon the following set of achievements: the agent architecture proposed in this article, the model of a category with a prototype, the strategy for learning categories with prototypes, the definition of cognitive semantics, and the analytical verification of the cognitive semantics' properties.

The proposed architecture model lists the necessary modules and describes their functionality, at the same time indicating their importance for the implementation of the main pragmatic goal assigned to the agent—namely, the commonsense acceptable (logically and pragmatically consistent) generation of autoepistemic membership statements. In this sense, the proposed architecture model can further serve as a reference point for research involving intentional conceptualization and intentional explanation of the meaning of individual stages of the cognitive process responsible for complete handling of the production of the considered cases of autoepistemic statements.

Author Contributions: Conceptualization, R.P.K.; methodology, M.Ż.; validation, R.P.K., G.P. and M.Ż.; formal analysis, M.Ż.; investigation, M.Ż.; resources, M.Ż.; data curation, M.Ż.; writing—original draft preparation, M.Ż. and G.P.; writing—review and editing, R.P.K., G.P. and M.Ż.; visualization, R.P.K., G.P. and M.Ż.; supervision, R.P.K. and G.P.; project administration, R.P.K. and G.P.; funding acquisition, R.P.K. and G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ogden, C.K.; Richards, I.A. *The Meaning of Meaning*; Harcourt, Brace: New York, NY, USA, 1923.
2. Eco, U. *La Struttura Assente*; Bompiani: Milan, Italy, 1968.
3. Vogt, P. The Physical Symbol Grounding Problem. *Cogn. Syst. Res.* **2001**, *3*, 429–457. [CrossRef]
4. Vogt, P. Language evolution and robotics: Issues on symbol grounding and language acquisition. In *Artificial Cognition Systems*; IGI Global: Hershey, PA, USA, 2006; pp. 176–209.
5. Steels, L. Language games for autonomous robots. *IEEE Intell. Syst.* **2001**, *16*, 16–22. [CrossRef]
6. Steels, L. Agent-based models for the emergence and evolution of grammar. *Philos. Trans. R. Soc. Biol. Sci.* **2016**, *371*, 20150447. [CrossRef] [PubMed]
7. Talmy, L. *Toward a Cognitive Semantics: Volume 1: Concept Structuring Systems and Volume 2: Typology and Process in Concept Structuring*; Bradford Book; MIT Press: Cambridge, MA, USA, 2003.
8. Wierzbicka, A. Conceptual primes in human languages and their analogues in animal communication and cognition. *Lang. Sci.* **2004**, *26*, 413–441. [CrossRef]
9. Kleiber, G. *La Semantique du Prototype. Categories et Sens Lexical*; Presses Universitaires de France: Paris, France, 1990.
10. Langacker, R.W. *Foundations of Cognitive Grammar*; Stanford University Press: Redwood City, CA, USA, 1987; Volume 1.
11. Lakoff, G. Categories. In *Proceedings of the Linguistics in the Morning Calm*; Hanshin: Seoul, Republic of Korea, 1982.
12. Geeraerts, D. On Necessary And Sufficient Conditions. *J. Semant.* **1986**, *5*, 275–291. [CrossRef]
13. Rosch, E.H. Natural categories. *Cogn. Psychol.* **1973**, *4*, 328–350. [CrossRef]
14. Rosch, E.H.; Mervis, C.B. Family resemblances: Studies in the internal structure of categories. *Cogn. Psychol.* **1975**, *7*, 573–605. [CrossRef]
15. Wang, B.; Yang, K.; Zhao, Y.; Long, T.; Li, X. Prototype-Based Intent Perception. *IEEE Trans. Multimed.* **2023**, *25*, 8308–8319. [CrossRef]
16. Du, Y.; Zhou, D.; Xie, Y.; Lei, Y.; Shi, J. Prototype-Guided Feature Learning for Unsupervised Domain Adaptation. *Pattern Recognit.* **2023**, *135*, 109154. [CrossRef]
17. Zhou, L.; Li, N.; Ye, M.; Zhu, X.; Tang, S. Source-free domain adaptation with Class Prototype Discovery. *Pattern Recognit.* **2024**, *145*, 109974. [CrossRef]
18. Katarzyniak, R. On some properties of grounding simple modalities. *Syst. Sci.* **2005**, *31*, 59–86.
19. Katarzyniak, R. On some properties of grounding nonuniform sets of modal conjunctions. *Int. J. Appl. Math. Comput. Sci.* **2006**, *16*, 399–412.
20. Katarzyniak, R. On some properties of grounding uniform sets of modal conjunctions. *J. Intell. Fuzzy Syst.* **2006**, *17*, 209–218.
21. Katarzyniak, R. Some notes on grounding singletons of modal conjunctions. *Syst. Sci.* **2006**, *32*, 45–55.
22. Katarzyniak, R.; Pieczyńska-Kuchtiak, A. A consensus based algorithm for grounding belief formulas in internally stored perceptions. *Neural Netw. World* **2002**, *12*, 461–472.
23. Katarzyniak, R.; Prusiewicz, A. Grounding and extracting modal responses in cognitive agents: 'and' query and states of incomplete knowledge. *Int. J. Appl. Math. Comput. Sci.* **2004**, *14*, 249–263.
24. Skorupa, G.; Katarzyniak, R. Applying Possibility and Belief Operators to Conditional Statements. In *Proceedings of the Knowledge-Based and Intelligent Information and Engineering Systems*; KES 2010, Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6276, pp. 271–280. [CrossRef]
25. Skorupa, G.; Katarzyniak, R. Conditional Statements Grounded in Past, Present and Future. In *Proceedings of the Computational Collective Intelligence, Technologies and Applications*; ICCCI 2010, Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6423, pp. 112–121. [CrossRef]
26. Katarzyniak, R.; Lorkiewicz, W.A.; Więcek, D.P. Some notes on extracting linguistic summaries built with epistemic modalities and natural language connectives of equivalence. *Comput. Methods Sci. Technol.* **2017**, *23*, 19–41. [CrossRef]
27. Katarzyniak, R.; Więcek, D.P. A note on nonemptiness of cognitive semantics for linguistic representations of modal equivalence. *Comput. Methods Sci. Technol.* **2018**, *24*, 301–315. [CrossRef]
28. de Saussure, F. *Course in General Linguistics*; Philosophical Library: New York, NY, USA, 1959.
29. Langacker, R. *Cognitive Grammar: A Basic Introduction*; Oxford University Press: Oxford, UK, 2008. [CrossRef]
30. Langacker, R.W. Levels of Reality. *Languages* **2019**, *4*, 22. [CrossRef]
31. Freeman, W.J. Comparison of brain models for active vs. passive perception. *Inf. Sci.* **1999**, *116*, 97–107. [CrossRef]
32. Freeman, W.J. A neurobiological interpretation of semiotics: meaning, representation, and information. *Inf. Sci.* **2000**, *124*, 93–102. [CrossRef]

33. Paivio, A. *Mental Representations: A Dual Coding Approach*; Oxford Psychology Series; Oxford University Press: Oxford, UK, 1986.
34. Stacewicz, P.; Włodarczyk, A. To Know we Need to Share—Information in the Context of Interactive Acquisition of Knowledge. *Procedia Comput. Sci.* **2020**, *176*, 3810–3819. [CrossRef]
35. Włodarczyk, A. Roles and Anchors of Semantic Situations. In *Meta-Informative Centering in Utterances (Between Semantics and Pragmatics)*; John Benjamins: Amsterdam, The Netherlands, 2013; pp. 3–20. [CrossRef]
36. Włodarczyk, A.; Włodarczyk, H. Agents, roles and other things we talk about: Associative Semantics and Meta-Informative Centering Theory. *Intercult. Pragmat.* **2008**, *20*, 345–365. [CrossRef]
37. Włodarczyk, A. Grounding of the Meta-Informative Status of Utterances. In *Meta-Informative Centering in Utterances (Between Semantics and Pragmatics)*; John Benjamins: Amsterdam, The Netherlands, 2013; pp. 41–58. [CrossRef]
38. Włodarczyk, A.; Włodarczyk, H. Qu'est-ce au juste que la prédication ? *Bull. Soc. Linguist.* **2019**, *114*, 1–54.
39. Stachowiak, F.J. Tracing the Role of Memory and Attention for the Meta-Informative Validation of Utterances. In *Meta-Informative Centering of Utterances between Semantics and Pragmatics*; Wlodarczyk, H., Wlodarczyk, A., Eds.; John Benjamins: Amsterdam, The Netherlands, 2013; pp. 121–142.
40. Harnad, S. The symbol grounding problem. *Phys. Nonlinear Phenom.* **1990**, *42*, 335–346. [CrossRef]
41. Harnad, S. Computation Is Just Interpretable Symbol Manipulation: Cognition Isn't. *Minds Mach.* **1994**, *4*, 379–390. [CrossRef]
42. Vogt, P. Anchoring of semiotic symbols. *Robot. Auton. Syst.* **2003**, *43*, 109–120. [CrossRef]
43. Lorkiewicz, W.; Katarzyniak, R. Multi-participant Interaction in Multi-agent Naming Game. *Comput. Methods Sci. Technol.* **2014**, *20*, 59–60. [CrossRef]
44. Lipowska, D.; Lipowski, A. Emergence and evolution of language in multi-agent systems. *Lingua* **2022**, *272*, 103331. [CrossRef]
45. Gärdenfors, P. *Conceptual Spaces: The Geometry of Thought*; Bradford Book; MIT Press: Cambridge, MA, USA, 2000. [CrossRef]
46. Gärdenfors, P. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*; The MIT Press: Cambridge, MA, USA, 2014. [CrossRef]
47. Kriegeskorte, N.; Kievit, R.A. Representational geometry: Integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **2013**, *17*, 401–412. [CrossRef]
48. Hintikka, J. *Knowledge and Belief*; Cornell University Press: Ithaca, NY, USA, 1962.
49. Moore, R.C. Semantical considerations on nonmonotonic logic. *Artif. Intell.* **1985**, *25*, 75–94. [CrossRef]
50. Marek, V.W.; Truszczynski, M. Autoepistemic logic. *J. ACM* **1991**, *38*, 588–619. [CrossRef]
51. Przymusinski, T.C. Autoepistemic logic of knowledge and beliefs. *Artif. Intell.* **1997**, *95*, 115–154. [CrossRef]
52. Kripke, S.A. Semantical Analysis of Modal Logic I: Normal Modal Propositional Calculi. *Z. Math. Log. Grund. Math* **1963**, *9*, 67–96. [CrossRef]
53. Cohen, P.R.; Levesque, H.J. Intention is choice with commitment. *Artif. Intell.* **1990**, *42*, 213–261. [CrossRef]
54. Halpern, J.Y.; Moses, Y. Knowledge and Common Knowledge in a Distributed Environment. *J. ACM* **1990**, *37*, 549–587. [CrossRef]
55. Rao, A.; Georgeff, M. Modeling rational agents within a BDI-architecture. In *KR'91: Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1991; pp. 473–484.
56. Halpern, J.Y.; Moses, Y. A guide to completeness and complexity for modal logics of knowledge and belief. *Artif. Intell.* **1992**, *54*, 319–379. [CrossRef]
57. Singh, M.P. *Multiagent Systems: A Theoretical Framework for Intentions, Know-How, and Communications*; Springe: Berlin/Heidelberg, Germany, 1994.
58. Grosz, B.J.; Kraus, S. Collaborative plans for complex group action. *Artif. Intell.* **1996**, *86*, 269–357. [CrossRef]
59. van Lindern, B.; van der Hoek, W.; Meyer, J.J.C. Formalising abilities and opportunities of Agents. *Fundam. Inform.* **1998**, *34*, 53–101. [CrossRef]
60. Katarzyniak, R.; Popek, G.; Mulka, M.; Żurawski, M. Towards communicative agents with cognitive semantics of modal class-membership statements. In Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Changsha, China, 13–15 August 2016; pp. 1324–1329. [CrossRef]
61. Pawlak, Z. Information systems theoretical foundations. *Inf. Syst.* **1981**, *6*, 205–218. [CrossRef]
62. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
63. Dennett, D.C. True Believers:The Intentional Strategy and Why It Works. In *Mind Design II: Philosophy, Psychology, and Artificial Intelligence*; The MIT Press: Cambridge, MA, USA, 1997. [CrossRef]
64. Katarzyniak, R.; Popek, G.; Żurawski, M. Extracting categories with prototypes in artificial cognitive agents. *Procedia Comput. Sci.* **2020**, *176*, 3283–3292. [CrossRef]
65. Nguyen, N.T. *Advanced Methods for Inconsistent Knowledge Management*; Springer: Berlin/Heidelberg, Germany, 2008. [CrossRef]
66. High-Level Expert Group on AI. In *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
67. Xu, X.; Wang, Z.; Chi, Z.; Yang, H.; Du, W. Complementary features based prototype self-updating for few-shot learning. *Expert Syst. Appl.* **2023**, *214*, 119067. [CrossRef]