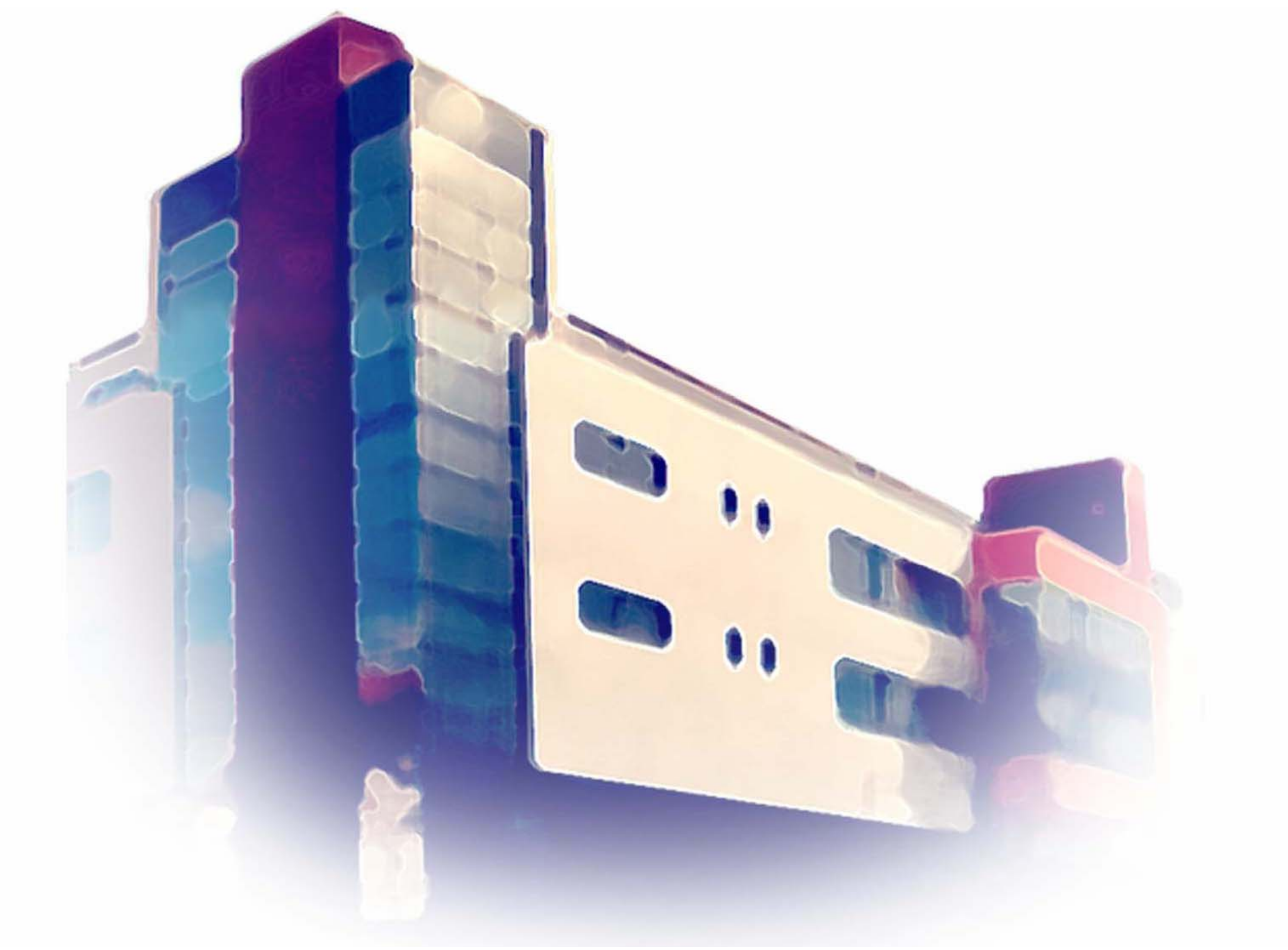
Lena Maria Kurzen

Analogy and Bisimulation: A Comparison of Indurkhya's Cognitive Models and Heuristic-Driven Theory Projection



PICS

Publications of the Institute of Cognitive Science

Volume 10-2005

ISSN: 1610-5389

Series title: PICS
Publications of the Institute of Cognitive Science

Volume: 10-2005

Place of publication: Osnabrück, Germany

Date: December 2005

Editors: Kai-Uwe Kühnberger
Peter König
Petra Ludewig

Cover design: Thorsten Hinrichs

UNIVERSITÄT OSNABRÜCK

Bachelor's Thesis

ANALOGY AND BISIMULATION: A COMPARISON OF INDURKHYA'S COGNITIVE MODELS AND HEURISTIC-DRIVEN THEORY PROJECTION

Lena Maria Kurzen
Cognitive Science, University of Osnabrück
lkurzen@uni-osnabrueck.de

Supervisors
Prof. Dr. Kai-Uwe Kühnberger
Dr. habil. Helmar Gust

August 2005

Abstract

The majority of formal approaches to analogy models analogies as structure preserving mappings from source to target domain. Bipin Indurkhya developed a formal theory of analogies and cognitive processes in general, where source and target are represented as concept networks, which are special kinds of algebras. The relation between them is modeled as an algebra as well. Another approach, which is also algebraic in nature, is heuristic-driven theory projection. It represents source and target domains as theories. Heuristic-driven theory projection is based on the idea that anti-unification can be used for modeling analogies. The analogical relation between source and target is induced by a heuristic algorithm that computes a third theory that is a generalization of source and target theories. In this thesis, the concepts of analogy and bisimulation, a notion expressing structural identity without being isomorphic, are brought together. The relations that model the connection between source and target in Indurkhya's theory and in heuristic-driven theory projection are compared to each other and to bisimulation. Conditions are specified under which these relations are bisimulations.

Acknowledgements

First and foremost, I would like to thank my supervisor Kai-Uwe Kühnberger for his valuable advice and support. Furthermore, I thank Simon Abbott, Stefan Bruda and Torsten Dierkes for helpful discussions and extensive proofreading.

Contents

1 Introduction

The interest in analogies is relatively high in cognitive science because analogies play a central role in various cognitive mechanisms. In its linguistic form, analogy is the explicit mentioning of relational likeliness of distinct situations [16]. That analogies can already be found in the earliest preserved literature, written about four thousand years ago, shows its importance. In the epic *Gilgamesh*, it says:

> ... Gilgamesh covered Enkidu's face with a veil like the veil of a bride. He hovered like an eagle over the body, or as a lioness does over her brood.

Analogy also refers to more general processes of 'seeing something as something else' [16]. Much work has been done to investigate the underlying principles of analogical reasoning. One of the fundamental empirical results in this field is that the interpretation of analogies and metaphors is mainly guided by the relational structures within the domains. This is the basic idea of Dedre Gentner's *structure-mapping theory* [13], an influential work, which most of the computational models of analogy are based on.

Bipin Indurkhya developed a formal theory for the modeling of metaphors, analogies and cognitive processes in general [25]. According to this theory, the processing of analogies is based on the same principles as perceptual processes. In this framework the domains of analogies are modeled as concept networks, which are special kinds of algebras. The cognitive relation between them is modeled as an algebra as well. Indurkhya specifies a property that some cognitive relations have, called *coherency*; coherent cognitive relations are structure preserving.

Heuristic-driven theory projection (HDTP) [20] is an approach that can model analogies and metaphors in domains that are more complex and require full first-order logic for representation. In HDTP the domains of an analogy are represented as theories in first-order languages. HDTP is based on the assumption that anti-unification can be used in models of analogy [19].

Given source and target theories, a heuristic algorithm computes a third theory that subsumes the input theories. This generalized theory is a structural description of both source and target. Additionally, facts and laws are transferred from source to target and thereby the creative aspects of analogy are modeled. The *analogical* *relation* between source and target relates corresponding items in source and target that have been generalized.

In both approaches, the one developed by Indurkhya and HDTP, the central relations between source and target ensure structural similarities of both domains. Bisimulation is a notion that expresses structural identity without being isomorphic. It is used in different fields, e.g. concurrent processes, game theory, modal logic and coalgebra. The question that arises here is how far the coherent cognitive relation in Indurkhya's theory and the analogical relation in HDTP differ from each other and whether they are bisimulations. In this thesis, this is investigated on a formal level, and coherent cognitive relations and analogical relations are compared to each other and to bisimulations. Precise conditions are specified for coherent cognitive relations and analogical relations for being bisimulations.

The remainder of this thesis is structured as follows: Section 2 will give an outline of some of the work that has been done to model human analogical reasoning. In Section 3, the basic ideas of Indurkhya's theory and heuristic-driven theory projection will be presented, concentrating on the formal concepts that are used to represent the entities involved in an analogy and the relations between them. Section 4 will give an overview of different variations of bisimulation that are used to express structural equivalences in various domains. In Section 5, bisimulation and the formal concepts of Indurkhya's cognitive models and heuristic-driven theory projection will be brought together. The structure preserving relations in both approaches will be compared to the notion of bisimulation. In Section 6, I will conclude this thesis and present an outline of future work.

2 Analogies - An Overview

2.1 Analogy

Different types of analogies can be distinguished [25]. Two of them are:

- Proportional analogy
  A proportional analogy can be formalized as $A : B :: C : D$ ('A is to B as C is to D'). A proportional analogy problem has the same underlying structure. Items A, B and C are given and item D has to be found. Such tasks are known from intelligence tests, where they occur in various domains, e.g. geometrical objects, words or numbers. An example of a proportional analogy is: $\square : \blacksquare :: \lozenge : \blacklozenge$.

- Predictive analogy
  A predictive analogy is used to explain a new domain by specifying similarities to a known one. A well-known example is the Rutherford analogy: *The hydrogen atom is like the solar system*, which can be used to explain Rutherford's atom model to someone who does not know it but has a conceptualization of the solar system.

Note that the Rutherford analogy can also be understood as a proportional analogy, formally represented as *sun : planet(s):: nucleus : electron*.
Analogy is not limited to the linguistic level but also refers to more general processes of 'seeing something as something else' and is one of the core processes of cognition. Analogy as a cognitive mechanism in general can be seen as 'the ability to think about relational patterns' [16].
Another related field is *analogical problem solving*, where the solution of a known problem is transferred to an unsolved problem in order to derive its solution. This is a very natural way for humans to solve problems.

Much work has been done in cognitive science to study the underlying processes of analogical reasoning. Analogy can be formalized as the establishment of a relation between two entities that are brought together in the analogy. One entity (referred to as 'target') is described by comparing it to another one ('source' or 'base'). The source is usually thought of to be better known than the target.

2.2 Analogy as Structure-Mapping

Psychology has investigated the processes of human analogical thinking. One of the major findings was that relational patterns, i.e. interrelationships between facts, play a central role in human analogical reasoning. This is one of the basic ideas of Dedre Gentner's *structure-mapping theory* [13], an influential work in the field of modeling analogical reasoning, which is based on the results of psychological experiments. According to this theory, the interpretation of analogies is guided by the relational structures within the domains rather than the attributes of single objects. In the example *'The hydrogen atom is like the solar system'*, the psychologically preferred interpretation tells us that the electron can be thought of as revolving around the nucleus because the interrelationship of the forces between nucleus and electron is similar to that between sun and planets in the solar system. An interpretation of the analogy that results in inferring that the nucleus is hot and yellow is clearly not preferred. What is important for the interpretation is the relation between sun and planets, and the one between nucleus and electron. When interpreting the analogy in the standard way, attributes like *hot(sun)* and *yellow(sun)* are not relevant.

According to the structure-mapping theory, an analogy can be formalized as a structure preserving mapping from source to target domain. Relations between objects of the source are transferred to the target domain and hold between the corresponding objects in the target. Analogous objects do not have to resemble each other but can have different attributes. What makes them analogous is that they play equivalent roles in relational structures. Gentner also formulated the *systematicity principle* [13], which says that mappings of larger structures containing higher-order relations, i.e. relations that take other relations as arguments, are preferred. This principle formalizes what has been found out in empirical studies: Humans have a tendency towards interpretations of analogical mappings that preserve the in-depth structure of the source domain [13].
The majority of computational models of analogy follow the ideas formulated in Gentner's structure-mapping theory. They use symbolic descriptions of source and target that allow for explicit representation of relations; most approaches use variations of predicate calculus to represent the two analogous domains.

2.2.1 The Structure-Mapping Engine

An implementation based on the structure-mapping theory is the *structure-mapping engine* (SME) [7, 8], a program simulating human analogical reasoning. As input SME is provided with the propositional descriptions of source and target domain in predicate/argument structure. The central part of SME's procedure is the establishment of syntactically consistent analogical mappings from source to target. The mappings are defined not only on the objects of the domains, but also predicates (except one-place predicates) associated with an object are mapped. This reflects what is explained above: The interpretation of analogies by humans is guided by the relational structures of the domains rather than the attributes of the objects. Which of the relations in the source are actually mapped is determined by the systematicity principle. Relations that are parts of structures containing higher-order relations are preferred over relations that occur in isolation. SME has been tested on several examples; one of them is the Rutherford analogy. In this example, SME is provided with the following input:
The description of the source domain:
  yellow(sun),
  greater(temperature(sun), temperature(planet)),
  cause(attracts(sun, planet) and greater(mass(sun), mass(planet)),
    revolves-around(planet, sun)).
The description of the target:
  attracts(nucleus, electron),
  greater(mass(nucleus), mass(electron)),
  revolves-around(electron, nucleus).

Given these descriptions, the preferred mapping that is computed by SME maps *sun* to *nucleus* and *planet* to *electron* because this results in the highest structural agreement of source and target. Since the relations holding in the source are transferred to the target, one derives the causal explanation for why the electron revolves around the nucleus: cause(attracts(nucleus, electron) and greater(mass(nucleus), mass(electron)), revolves-around(electron, nucleus)).

As argued in [9], one constraint on cognitive simulations is the *integration constraint*: For models of analogies it means that a computational model of analogy should also be usable as a subpart of larger models of more general cognitive processes that involve processes of analogy. This condition is reasonable because cognition consists of various integrated subprocesses. Therefore, for an appropriate model of cognition one must strive for integrability of the different submodels.

2.2.2 MAC/FAC

The integrability of SME has been tested by integrating it into a system called MAC/FAC [15] that models similarity-based retrieval. Similarity-based retrieval occurs in situations where something in our environment reminds us of something we have encountered in the past. The similarity of both items can be on different levels: It can be structural, e.g. hearing octaves can remind one of the periodic tables in chemistry, or purely superficial, e.g. seeing someone who is wearing round glasses can remind one of the wheels of a bicycle [15].
The MAC/FAC program gets two inputs: a pool of items stored in the memory, representing the content of the long-term memory, and a probe, representing the new item. What MAC/FAC does is choosing one entity from its memory that best fits the probe, i.e. it selects the item of which one would most likely be reminded upon encountering the probe item. MAC/FAC models similarity-based retrieval as a two-stage process. First, possible candidates are chosen from the memory and in a second stage these candidates are evaluated and the most promising one is selected. In the first stage (MAC, "many are candidates"), a matcher is used that selects the candidates from the memory that will most likely give best results in the second stage where the best of the candidates is selected.
In the second stage (FAC, "few are chosen"), the candidates chosen in the first stage are evaluated. In this process, SME is used; correspondences between the candidates and the probe are computed and candidate inferences are added. Information from the memory item is transferred to the probe. The correspondences are evaluated according to the level of similarity between candidate and probe. Finally, the candidate with the best results is chosen.

2.2.3 Phineas

SME has also been used as a subpart of *Phineas* [6], a system that models the learning of theories in the domain of physics. Phineas 'learns' physical theories by analogy with previously understood examples. It has a database where known physical theories are stored, e.g. theories explaining boiling, osmosis, liquid flow and oscillation [6, 9]. As input Phineas gets the qualitative description of the behavior of a physical system. In one of the examples, Phineas has been tested on, it is provided with a description of the temperature changes that occur when a hot brick is immersed in cold water. Given this description, Phineas first tries to 'understand' the behavior by applying one of the physical theories of its database to it. In this stage, the qualitative simulator *QPE* is used, which produces predictions about the possible behaviors of a system in a certain situation that are consistent with the qualitative laws of a given theory.
Using QPE, Phineas simulates the possible behaviors of the system on the base of the theories in its database. In the next step, Phineas uses a measurement interpretation

system to construct explanations of the behavior of the system under consideration. This measurement interpretation system explains the data the of a system's behavior by comparing the data to the simulations based on the known theories. If in the example mentioned above, Phineas does not have a theory of heat flow in its database, it tries to explain the temperature changes of water and brick with an analogous theory; it searches in its database of previously explained examples. This is the point where SME comes into play since the previously understood example with the highest structural similarity to the water-brick-situation has to be found because that is the one that will probably explain the current situation in the best way. In the example, Phineas chooses 'water flow' from its memory and applies this theory to the example of the brick in the water. Thus, the concept of 'water flow' is used to explain the 'heat flow' in the situation with the hot brick in the cold water.

It has been criticized that the models presented above, wich are based on SME, do not take into account the limited capacity of human working memory [24]. A huge amount of data is processed in parallel in these programs. In this respect, they do not model the underlying processes of human analogical thinking in an appropriate way. Holyoak and Hummel introduce an alternative architecture for analogical reasoning. Their system LISA [23] seems to be closer to psychological and neuroscientific findings. Their idea is that not only the symbolic nature of cognitive processes has to be modeled but the model should also be embedded in a neurally inspired environment because that is where the processes take place in humans.

2.3 Analogy as High-Level Perception – Copycat

Another point of criticism of SME and other similar models is that the representations of source and target are explicitly given and kept fixed. It is clear that the analogical mapping depends crucially on how source and target are represented. This argument can be seen as a starting point for the approach by Falkenhainer and Mitchell who developed a system called *Copycat* [22]. Copycat operates in a microworld of alphabetical strings. The following example illustrates the analogical problems in Copycat's microworld [22]:

If *abc* is changed to *abd*, how would you change *yk* in the same way?

'In the same way' has to be defined more precisely but empirical studies revealed that the most frequent answer is *yl*, which seems to be the natural way to solve the problem. The underlying rule that is used here is: *Replace the rightmost character by its alphabetic successor*. Now, consider a slightly modified version of the problem:

If *aabc* is changed to *aabd*, how would you change *ykk* in the same way?

Here, the answer does not seem as obvious as the one in the previous problem. One possibility to solve this problem is clearly to proceed in the same way as in the first one, i.e. the rightmost character is again replaced by its successor which yields the result *ykl*. But for most people there is another way to solve it that seems more appropriate. In contrast to the first problem, the *a* and the *k* are doubled, which changes the appearance of the strings: The two *a*s and *k*s are perceived as groups. Hence, changing *ykk* to *yll* seems more natural than transforming it into *ykl*. This phenomenon is referred to as *mental fluidity*. It means that one concept can slip into another one because the 'distances' between concepts change with the evolving perception. In the example above, the concept *rightmost character* slips into the concept *rightmost group*. The microworld is not designed to model only analogical reasoning in the domain of alphabetic strings but it is supposed to stand for other domains closer to reality. A concept like *successor of* can stand for other relational concepts like *brother of* or *neighbor of*. The phenomenon of mental fluidity is reflected in the *slipnet*, a central part of Copycat's architecture, which contains concept types whose distances change because they are influenced by the process of perception.

Although the the approaches presented in this section differ in many aspects, they all focus on the relations that hold between the entities in the analogous domains.

3 Algebraic Approaches to Analogy

In this section, two approaches to analogy will be presented that differ from other approaches in that they have formal specifications and do not work primarily example-based as the ones presented in the previous section. According to the structure-mapping theory, analogy can be formalized as a structure preserving mapping from source to target. This idea is also respected in the following approaches, which are both based on algebraic frameworks.

3.1 Indurkhya's Theory

3.1.1 Formal Concepts

The theory developed by Bipin Indurkhya [25] is not only an approach to the modeling of metaphors and analogies but proposes a model for cognitive processes in general. I will focus on the formal concepts of this approach and how analogies can be modeled in this framework, and not on the philosophical basis of Indurkhya's view of cognition in general. He proposes a formal algebraic framework for modeling cognitive processes. In some respects, the underlying ideas of this approach are similar to the ones of Falkenhainer's Copycat because both emphasize the connection of perception and higher cognitive processes. According to them, higher cognitive processes have the same underlying principles as perceptual processes.

In Indurkhya's framework, perceptual processes are modeled as an interrelation of a *source concept network*, which stands for the conceptualization the agent has of reality, and the sensorimotor data the agent has of it (referred to as 'environment') [25]. The concept network and the system of sensorimotor information are represented as *algebras*.

Definition 1 (Indurkhya)**.** *An* ***algebra*** *is a pair* $\langle A, \Omega \rangle$*, where* A *is a non-empty set of objects and* Ω *is a set of operators. Associated with each operator is its arity, i.e. the number of arguments it takes. An operator of arity* n *is a function* $\omega : A^n \rightarrow A$*.* $\Omega(n)$ *denotes the set of operators of arity* n*.*

The relation between both algebras, the connection between source concept network and environment, is also modeled as an algebra. This results in three algebras, which together form a so called *cognitive model*. Let us consider these three components in more detail:

Definition 2.

- *The* ***source concept network*** $\mathscr{A} = \langle A, \Omega \rangle$ *is an algebra satisfying the following conditions:*

    1. *The class of operators* Ω *is either finite or a subset of the polynomial operations generated by a finite subset of* Ω.
    2. *Every operator is a computable function.*
    3. *The algebra is finitely generated, i.e. there is a finite set* $X \subseteq A$ *such that every element of* A *can be generated from elements of* X *by using the operators of* Ω.

- *The* ***environment*** *is represented as an algebra* $\mathscr{B} = \langle B, \Sigma \rangle$.

- *The relation between source concept network and environment, referred to as* ***cognitive relation****, is an algebra* $\mathscr{R} = \langle R, \Psi \rangle$ *with* $R \subseteq A \times B$ *and* $\Psi(n) \subseteq \Omega(n) \times \Sigma(n)$ *for all arities* n *and*

    1. $\Psi^{-1}(\sigma) \neq \emptyset$ *for all* $\sigma \in \Sigma$.
    2. $\langle R, \Psi \rangle^{-1}(\langle B, \Sigma \rangle) \subseteq \langle A, \Omega \rangle$ *is a finitely generated concept network.*

- $\mathcal{C} = \langle \langle A, \Omega \rangle, \langle R, \Psi \rangle, \langle B, \Sigma \rangle \rangle$ *is called* ***cognitive model****.*

Note that the constraints for the source algebra are stronger than the ones for the environment algebra. This is due to the fact that the sensorimotor data one has of reality can itself be seen as more or less unstructured, and structural information is obtained when it is interpreted in terms of the conceptualizations one has of the environment. It is the source concept network that induces the structure of the environment via the cognitive relation. This can be seen in the first condition for the cognitive relation, which says that each operator of the environment is related to an operator of the source concept network. If there were transitions in the environment which did not correspond to any operators in the source concept network, the cognitive agent would not have a conceptualization of these transitions. Condition 1 ensures a certain degree of structural overlap of concept network and environment on the operator level. The second condition says that the subset of the source concept network that is related to the environment is itself a finitely generated concept network. Thus, finite representability of the cognitive model can be ensured.

An important property of some cognitive relations is that they preserve the operations in the source concept network and the transformations in the environment. This property is called *coherency*. It is quite clear that coherency is an ideal a cognitive agent must strive for in order to make sense out of its sensorimotor data of the environment. This idea of a structure preserving cognitive relation is similar to the properties of the analogical mapping in Gentner's structure-mapping theory, which is structure preserving as well.
That leads us to the notion of coherency, which is investigated in more detail in Section 5.

Definition 3 (Indurkhya)**.** *Given a cognitive model $\mathcal{C} = \langle\langle A, \Omega\rangle, \langle R, \Psi\rangle, \langle B, \Sigma\rangle\rangle$ and a class $X \subseteq A$, we say that $\mathcal{C}$ (or $\langle R, \Psi\rangle$) is **locally coherent** in X if and only if whenever $x_1, \ldots, x_n \in X$, $\omega \in \Omega(n)$ and $\omega(x_1, \ldots, x_n) \in X$, then for any $y_1, \ldots y_n \in B$ and $\sigma \in \Sigma(n)$ with $\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle \in R$ and $\langle \omega, \sigma\rangle \in \Psi(n)$, it is the case that $\langle \omega(x_1, \ldots, x_n), \sigma(y_1, \ldots y_n)\rangle \in R$.*
*Conversely, $\mathcal{C}$ (or $\langle R, \Psi\rangle$) is called **locally coherent** in $Y \subseteq B$ if and only if whenever $y_1, \ldots, y_n \in Y$, $\sigma \in \Sigma(n)$ and $\sigma(y_1, \ldots, y_n) \in Y$, then for any $x_1, \ldots x_n \in A$ and $\omega \in \Omega(n)$ with $\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle \in R$ and $\langle \omega, \sigma\rangle \in \Psi(n)$, it is the case that $\langle \omega(x_1, \ldots, x_n), \sigma(y_1, \ldots y_n)\rangle \in R$.*

Definition 4 (Indurkhya)**.** *A cognitive relation $\langle R, \Psi\rangle$ of a cognitive model $\mathcal{C} = \langle\langle A, \Omega\rangle, \langle R, \Psi\rangle, \langle B, \Sigma\rangle\rangle$ is called **coherent** if it is locally coherent in A and B.*

Note that $\langle R, \Psi\rangle$ is locally coherent in A if and only if it is locally coherent in B.
In order to be able to specify further properties of cognitive models, Indurkhya introduces several terms that characterize them.

Definition 5 (Indurkhya)**.** *Let $\mathcal{C} = \langle\langle A, \Omega\rangle, \langle R, \Psi\rangle, \langle B, \Sigma\rangle\rangle$ be a cognitive model.*

1. *$x \in \langle A, \Omega\rangle$ is called **relevant** if $\langle R, \Psi\rangle(x) \neq \emptyset$. If every $x \in \langle A, \Omega\rangle$ is relevant, $\mathcal{C}$ is called **full**.*

2. *If $\langle R, \Psi\rangle(x_1) = \langle R, \Psi\rangle(x_2)$ for $x_1, x_2 \in \langle A, \Omega\rangle$, x_1 and x_2 are called **synonymous**. If there is no pair of synonymous distinct entities in $\langle A, \Omega\rangle$, $\mathcal{C}$ is called **optimal**.*

3. *An element y of the environment $\langle B, \Sigma\rangle$ is said to be **visible** if $\langle R, \Psi\rangle^{-1}(y) \neq \emptyset$. If every element of the environment is visible, the cognitive model is called **complete**.*

4. *If the grouping induced by the cognitive relation $\langle R, \Psi\rangle$ on the environment is pairwise disjoint, the cognitive model is called **unambiguous**.*

5. *Two items* $x, y \in \langle B, \Sigma \rangle$ *are* ***indistinguishable*** *if* $\langle R, \Psi \rangle^{-1}(x) = \langle R, \Psi \rangle^{-1}(y)$. *If there is no pair of indistinguishable elements,* $\mathcal{C}$ *is called* ***fully resolved****.*

Relevant entities of the source concept network are therefore all these that are used for interpreting the environment. Two synonymous entities are both related to the same set of objects or transitions of the environment. In an optimal cognitive model distinct elements of the source concept network are related to distinct sets of entities of the environment. An entity being visible means that the agent has a concept of it. In an unambiguous cognitive model, for each entity in the environment there is at most one corresponding one in the source concept network and therefore it never happens that there is more than one way of interpreting an entity of the environment. If the cognitive model is fully resolved the agent is able to distinguish between all the transitions and objects in the environment.

3.1.2 Modeling Analogies

After this overview of the central formal concepts of Indurkhya's theory, let us now consider how analogies can be modeled in this framework. Analogies and metaphors can roughly be described as processes of 'seeing something as something else' [16]. In Indurkhya's theory, such processes can be modeled as a non-standard interpretation of the environment [25].
The target is interpreted in terms of the concept network of the source item. In the case of the Rutherford analogy, this means that the atom system is interpreted using the concept network of the solar system. In general, such a process is similar to the perceptual processes described above; it just takes place at a higher level, i.e. as a relation between two concept networks. If we come back to the constraints for the cognitive model (cf. Definition 2), we can see that they make sense for analogies as well. In case of analogies, the first condition for the cognitive relation ensures that every operator of the target domain can be interpreted as an operator of the source domain. If this were not the case, the analogy would be perceived as inconsistent since there would be transitions in the target domain that do not have counterparts in the source and thus cannot be explained in terms of source operators.
The structural similarity of source and target, which is ensured by this condition, is a reasonable constraint for analogies as well because structural overlap of source and target makes it easier to understand the analogy[1]. The second condition, which says that the relevant part of the concept network is itself a finitely generated concept network, can be interpreted for the modeling of analogies as follows: One might have an immense amount of knowledge of the source domain under consideration but for an

[1]In [30], empirical evidence is given for the impact of structural overlap of source and target domain on the performance of humans in analogical problem solving tasks.

interpretation only a relatively small subset might be relevant. In order to interpret the Rutherford analogy in the intended way, the knowledge one might have of the material of the Jupiter surface is not relevant. The relevant subset of the source should be consistent; this corresponds to the fact of the source being a concept network.

Indurkhya's theory is applied to analogies in a character string domain, similar to the one Copycat is operating in [4, 5]. The objects in this domain are strings of characters and the operators represent different kinds of pattern regularities like iteration, symmetry, concatenation and alternation. Atomic elements are single characters or groups of characters that are perceived as indivisible units. The general idea of this model is similar to that of Falkenhainer's Copycat [22]: Given a proportional analogy problem in the string domain, i.e. given two strings A and B and a third string C, a string D has to be found such that $A : B :: C : D$. This is done by first finding appropriate representations of the given strings using the operators. Then these representations are evaluated by calculating their *information load* [37], which represents the complexity of the chosen representations. In the next step, the representations with the lowest information loads are chosen and the string D is computed by applying the rule that transforms A into B to C. Choosing representations with low information loads is related to the tendency of humans to prefer simple patterns over more complex ones. Consider the following example: The preferred perceptual pattern of *abccba* is seeing it as a symmetry structure made of the string *abc*, which itself is a successor structure of three items starting with *a*. An alternative would be seeing *abccba* as a symmetry pattern made of the string *ab* (a successor structure of two items staring with *a*) with the item *cc* in between. The second representation is less preferred because it is more complex.

3.2 Heuristic-Driven Theory Projection

Indurkhya's approach seems to be suitable only for domains that can be formalized in a rather simple algebraic way. *Heuristic-driven theory projection* (HDTP) [20] takes into account this deficiency of the former approach. A theory for modeling the interpretation of metaphoric expressions is proposed. It is based on an algebraic framework as well, but in contrast to the approaches mentioned so far, it represents source and target domain as theories, i.e. as sets of facts and laws, and can thereby model metaphors and analogies in domains that require full first-order logic for representation. Another deficiency of the previously described approaches is that usually it is presupposed that source and target are taken from the same domain. Thus, only a restricted subset of analogies can be modeled. In contrast, HDTP can be used to model analogies between two different domains that are encoded in two different theories [20].

3.2.1 The Underlying Assumptions

This approach builds upon the hypothesis that there is a strong similarity between some types of metaphors and analogies[2]. That this assumption is reasonable can be illustrated by the following example [20]:

1. Gills are the lungs of fish.

2. Gills are to fish as lungs are to mammals.

The proportional analogy (2) expresses the same meaning as the metaphor (1) but in a more explicit way. Because of these similarities to analogies, the interpretation of metaphoric expressions can be investigated formally with the same methods that are known from modeling analogical reasoning.
Another hypothesis, the approach presented in [20] builds on, is that there is more information given about the source than about the target. This corresponds to the idea that analogy (especially predictive analogy) can be seen as a process of applying knowledge of a known situation to a less familiar one.

3.2.2 Modeling the Domains

In HDTP source and target domain are modeled as sets of axioms, which induce respective theories.
If we come back to the Rutherford analogy and the SME model of it, one might have noticed that the modeling of the domains seems in some respect inappropriate. The *attracts*-relation in the target domain is not primarily based on the fact that the mass of the nucleus is greater than the mass of the electron, as it is modeled in SME, but it is rather caused by the difference in electric charge. This is not modeled in an appropriate way. Furthermore, the fact that the electron revolves around the nucleus is explicitly given in the target domain as it is modeled in SME. This seems counterintuitive in some cases where the analogy is used to explain Rutherford's atom model to someone who does not know it yet. In such a case, the fact that the electron revolves around the nucleus is rather thought to be what can be inferred from the analogy. It can be seen as a new fact that is transferred from the source and introduced to the target. One important feature of HDTP is that only known facts are given in the target. In case of analogies in naive physics, the initial representation of the target domain contains only facts that can be objectively measured or that

[2]Support of this hypothesis is given in [14].

can be found out experimentally. That the electron is revolving around the nucleus cannot be measured and is therefore not part of the representation of the initial target domain. The target domain, as it is modeled in HDTP, contains information about the differences in charge and mass of nucleus and electron and the fact that the distance between them is always greater than zero, which is the result of Rutherford's scattering experiment.

In contrast to Indurkhya's theory, in HDTP a many-sorted algebra is used for representation. Therefore, it can be represented explicitly that objects can be of different sorts and operators can have sortal restrictions.

Definition 6 (Gust, Kühnberger, Schmid)**.** *A* ***many-sorted signature*** Σ *is a set* $\Sigma = \{Sort_\Sigma, Func_\Sigma, Type_\Sigma\}$. *$Sort_\Sigma$ is a partially ordered set of sorts, $Func_\Sigma$ is a finite set of function symbols and $Type_\Sigma$ is a function, $Type_\Sigma : Func_\Sigma \rightarrow Cl(Sort_\Sigma)$, where $Cl(Sort_\Sigma)$ is the closure of sorts under products.*

Then a *term algebra* can be defined:

Definition 7 (Gust, Kühnberger, Schmid)**.** *Given a many-sorted signature Σ, the* ***term algebra*** *$Term(\Sigma, V, C)$ relative to a set of sorted variables $V = \{x_1 : s_1,\ x_2 : s_2, ...\}$ with $s_i \in Cl(Sort_\Sigma)$, and a finite set of constants $C = \{a_1 : s_1, \ldots, a_n : s_n\}$ with $s_j \in Cl(Sort_\Sigma)$, is defined as the smallest set such that the following conditions hold:*

1. *If $x : s \in V$ is given, then $x : s \in Term(\Sigma, V, C)$.*
2. *If $a : s \in C$ is given, then $a : s \in Term(\Sigma, V, C)$.*
3. *If $f \in Func_\Sigma, Type_\Sigma(f) = s_1 \times s_2 \times \cdots \times s_n \rightarrow s$ and $\forall i \in \{1, 2, \ldots, n\}$: $Type_\Sigma(t_i) = s_i$ is given, then $f(t_1, \ldots, t_n) \in Term(\Sigma, V, C)$ and $Type_\Sigma(f(t_1, \ldots, t_n)) = s$.*

The sets of terms of source and target are represented as term algebras, denoted by $Term_S$ and $Term_T$ respectively. For the representation of source and target domains, first-order predicate languages relative to the respective term algebras are defined.

Definition 8 (Gust, Kühnberger, Schmid)**.** *The* ***languages*** *L_S and L_T of source and target domain respectively are standard many-sorted first-order predicate logic languages relative to a given term algebra $Term(\Sigma, V, C)$. The following sub-languages can be specified (sortal restrictions are omitted in order to simplify readability):*

Terms $t := x \mid c \mid f^n(t_1, \ldots, t_n)$
Logical constants $l := \wedge \mid \vee \mid \rightarrow \mid \leftrightarrow \mid \neg \mid \forall \mid \exists \mid =$
Atomic (well-formed) formulas $\alpha := t = t' \mid R^n(t_1, \ldots, t_n)$
Well-formed formulas $\phi := \alpha \mid \phi \wedge \psi \mid \phi \vee \psi \mid \neg\phi \mid \phi \rightarrow \psi \mid \phi \leftrightarrow \psi \mid \forall x\phi \mid \exists x\phi$

Given these languages, the theories of source and target can be defined. They are given by a finite set of axioms that specify the facts that hold in the domains and the laws that can be used to derive new facts.

Definition 9 (Gust, Kühnberger, Schmid)**.** *A* ***theory*** *Th of a language L is specified as a consistent and finite set of well-formed formulas of L of the following form:*
Facts: $\alpha := \ R^n(t_1, \dots, t_n) \mid t = t' \mid \alpha \wedge \beta \mid \alpha \vee \beta$
Laws: $\phi \rightarrow \psi \mid \forall x \phi \mid \exists x \phi$

3.2.3 Establishing an Analogical Relation

Unlike the theory presented in [25], HDTP not only describes the properties of the analogical relation between source and target but also focuses on the underlying principles of the establishment of the relation.

An analogy expresses structural similarities of source and target domain. A generalization of both, source and target, or better the most specific generalization of them, captures exactly these structural similarities. Such a generalization can be found using *anti-unification*. In [19], it is shown that anti-unification can be used for modeling analogical reasoning. Anti-unification is the dual notion of *unification*, which is widely known (e.g. in Prolog) as a method for finding an instantiation of two terms. The result of a unification of two terms is a term that instantiates both of the input terms, or in other words, a term which is subsumed by the input terms.
As an example, consider the terms $F(X, b)$ and $F(a, Y)$ where F is a function symbol, X and Y are variables and the lower case letters are constants. Unifying them results in terms that instantiate both formulas and in corresponding substitutions. In this example, one solution is $F(a, b)$ with corresponding substitutions $X \mapsto a, Y \mapsto b$. Conversely, the result of an anti-unification of two terms is a third term that subsumes the two given ones together with corresponding substitutions that denote how the more specific terms can be generated from the anti-instance. Given the terms $F(a, b)$ and $F(X, c)$, a possible anti-instance is $F(X, Y)$ with substitutions $\theta_1/\theta_2 : F \mapsto F/F,\ X \mapsto a/X,\ Y \mapsto b/c$. Of course, when anti-unifying two terms, the trivial anti-instance Z (where Z is a variable not contained in the terms that are anti-unified) is always a solution. What is of interest, is a set of anti-instances that is most specific, minimal and complete because it carries maximal information about the structure both input terms have in common (cf. [20] and [39] for a detailed description of such sets of anti-instances). In case of the term algebras used in HDTP, a substitution on terms is a partial function mapping variables to terms: $\theta : V \rightarrow Term(\Sigma, V, C)$, $\theta = \{x_1 \mapsto t_1, \dots, x_n \mapsto t_n\}$, for $x_i \neq x_j$, if $i \neq j$. Substitutions are extended recursively to complex terms:

Definition 10 (Gust, Kühnberger, Schmid). *Given a term algebra $Term(\Sigma, V, C)$ and a substitution $\theta : V \rightarrow Term(\Sigma, V, C)$, a function sub is recursively defined:*

$sub(\theta, c) = c$ *if* c *is a constant*

$sub(\theta, v) = value(\theta, v)$ *if* $v \in dom(\theta)$

$sub(\theta, f(t_1, \ldots, t_n)) = f(sub(\theta, t_1), \ldots, sub(\theta, t_n))$

with $dom(\theta) = \{x | x \mapsto t \in \theta\}$

$range(\theta) = \{t | x \mapsto t \in \theta\}$

$value(\theta, v) = t$ *if* $v \mapsto t \in \theta$

For modeling analogies that involve whole theories, the method of anti-unification has to be extended such that not only terms but also complex formulas can be anti-unified. For finding a generalization of complex terms, first-order anti-unification is usually not sufficient but anti-unification of higher order is needed. In [20], it is shown that a subset of second-order problems can be specified that can be reduced to first-order problems[3]. For this purpose, equational theories E_S and E_T for source and target theories are introduced. They contain equations in solved form and thereby specify how terms can be replaced by equivalent ones. Source and target theories can be expanded relative to these equational theories. In order to reduce second-order operations to first-order ones, new function symbols are introduced, which can be used instead of the old ones, where the arguments are permuted or replaced by subterms. This expansion does not result in infinitely many computations because there is only a finite number of subterms (cf. [20] for a more detailed investigation of this argument). The resulting new terms can be used for anti-unification, and a first-order anti-unification of them is equivalent to a second-order anti-unification of the original ones. Using this method, a subset of second-order anti-unification can be reduced to first-order operations. This subset seems to be sufficient for the modeling of analogies and metaphors [20].
The establishment of an analogical relation between source and target theory is performed by the heuristic algorithm *HDTP-A*[4]. As input, it is provided with the axioms specifying source and target theories. Given these sets of axioms that induce source and target theories, Th_S and Th_T respectively, HDTP-A generates a set of axioms G that induces a theory Th_G, which is a generalization of source and target theories. HDTP-A also computes corresponding substitutions that specify how the axioms of source and target theory can be generated from the ones of Th_G. By-products of the algorithm are modified source and target theories $Th_S^{A_h}$ and $Th_T^{A_h}$ respectively.

[3] A similar idea is presented in [39] in the domain of analogical programming.

[4] The description given here is based on the implementation in [19].

HDTP-A can be sketched as follows (taken from [20]):

T = axioms of the target domain sorted by a heuristics h
S = axioms of the source domain
G = empty list of axioms of the generalized theory
$\theta_1 = \theta_2$ = empty substitutions
$Th_T^{A_h} = Th_T$
FOR $\psi \in T$
    $\psi = normal\_form(\psi)$
    SELECT $\phi \in S$
        $\phi = normal\_form(\phi)$
        IF not $same\_structure(\phi, \psi)$ REJECT
        SELECT$(\xi, \theta_1, \theta_2) \in anti_instances(\phi, \psi, \theta_1\theta_2)$
    WITH ξ best according to a heuristics h'
    IF $h'(\xi) >$ a given threshold
        ADD ξ to G
        ADD $\xi\theta_2$ to $Th_T^{A_h}$
        REMOVE ϕ from S
    ELSE FAIL
END FOR
FOR $\phi \in S$
    $\psi = transfer(\phi, \theta_1, \theta_2)$
    IF $T_T^{A_h} \vdash \neg\psi$ CONTINUE
    IF $oracle(\psi)$ = FALSE CONTINUE
    ADD ψ to $T_T^{A_h}$
    ADD $generalize(\phi, \theta_1)$ to G
END FOR

First, the input axioms are converted into a normal form (e.g. conjunctive normal form) because then it can be determined if they have equivalent structures. Then axioms from the target description are selected and matched to the ones of the source. The selection of target items is guided by a heuristics h. Appropriate heuristics, which can be used here, are ones that ensure that simple axioms are selected first[5]. Another heuristics, which reflects a similar idea as Gentner's systematicity principle, is one that says that axioms that maximize the number of shared terms with already generalized axioms are chosen first.
If a source axiom with identical structure is found, they are generalized. An anti-instance that is best according to a certain heuristics is chosen from the set of possible anti-instances. At this step of the algorithm, one can use heuristics that ensure that 'good' (i.e. cognitively preferred) anti-instances are chosen first. A possible heuristics that can be used is one which ensures that substitutions of minimal length are chosen first. It reflects the fact that humans usually prefer simple solutions over more complex ones. Another possibility is to use a heuristics saying that anti-instances with minimal number of second-order objects are preferred. It reflects a

[5]'simple' refers to a minimal number of embedded subterms and logical operators and a minimal arity of embedded relations.

similar idea as the first heuristics because second-order substitutions result in stronger structural changes than first-order ones and this leads to more complex solutions as well. After all target items have been anti-unified with corresponding source items, remaining source axioms that have not been anti-unified so far are transferred to the target. This is an important feature of HDTP because it models the creative aspects of analogy and the introduction of new concepts to the target domain. Before the actual transfer, it is checked whether transferring the axiom to the target theory would result in inconsistencies with the rest of the target. It is checked if the transferred axiom would lead to a contradiction in the theory and additionally an oracle is used, which represents an experiment. Only if the outcome of the experiment is consistent with the transferred version of the axiom, it is transferred to the target theory and a generalization is added to the the generalized theory. As many source axioms as possible are transferred to the target. This models the fact that interpretations with a large number of consistent correspondences between source and target are cognitively preferred.
This results in an enrichment of the target theory. On the source side, only facts and laws that have been anti-unified with corresponding items of the target are deleted from the initial set and added to $Th_S^{A_h}$, the resulting source theory (this is not represented in the pseudo code above). The source theory is thereby reduced to a subset that has corresponding counterparts in the target and therefore it represents the subset of the source that is actually used for interpreting the target.

The analogical relation between the modified source and target theories is defined as follows[6]:

Definition 11. *Given the theories $Th_S^{A_h}$ and $Th_T^{A_h}$ with corresponding models $\mathfrak{M}_S$ and $\mathfrak{M}_T$ and a coproduct operation (disjoint union) $\oplus$, an **analogical relation***

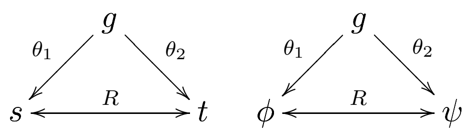
$$R \subseteq (Th_S^{A_h} \times Th_T^{A_h}) \oplus (Term_S \times Term_T)$$

of theories is a set of pairs $\langle x, y\rangle$ such that it holds that:

1. *If $\langle \phi, \psi\rangle$ is a pair of formulas, then $\langle \phi, \psi\rangle \in R$ if and only if there exists $g \in Th_G^{A_h}$ such that $(g, \{\theta_1, \theta_2\})$ is an anti-instance of ϕ and ψ and $Th_S^{A_h} \cup E_S \vdash g\theta_1 \leftrightarrow \phi$ and $Th_T^{A_h} \cup E_T \vdash g\theta_2 \leftrightarrow \psi$.*
2. *If $\langle s, t\rangle$ is a pair of terms, then $\langle s, t\rangle \in R$ if and only if there exists $g \in Th_G^{A_h}$ such that $(g, \{\theta_1, \theta_2\})$ is an anti-instance of s and t and $E_S \vdash g\theta_1 = s$ and $E_T \vdash g\theta_2 = t$.*
3. *There exists a model $\mathfrak{M}_S$ such that $\mathfrak{M}_S \models Th_S^{A_h}$ if and only if there exists a model $\mathfrak{M}_T$ such that $\mathfrak{M}_T \models Th_T^{A_h}$.*

[6]The definition, used here, is actually stronger than the one given in [20].

The syntactic constraints can be illustrated graphically:



The conditions can also be formulated in the following way: For terms $s \in Term_S$ and $t \in Term_T$, $\langle s, t\rangle \in R$ if and only if $E_S \vdash s = t\theta_2^{-1}\theta_1$ and $E_T \vdash t = s\theta_1^{-1}\theta_2$.

Analogously, for formulas $\phi \in Th_S^{A_h}$ and $\psi \in Th_T^{A_h}$, $\langle \phi, \psi\rangle \in R$ if and only if $Th_S^{A_h} \cup E_S \vdash \phi \leftrightarrow \psi\theta_2^{-1}\theta_1$ and $Th_T^{A_h} \cup E_T \vdash \psi \leftrightarrow \phi\theta_1^{-1}\theta_2$.

3.2.4 How HDTP Differs from other Approaches

As noted in [10], SME is only able to match the statements *”louder(Fred, Gina)”* and *”bigger(Bruno, Peewee)”* if they are represented as *”greater (loudness(Fred), loudness(Gina))”* and *”greater(size(Bruno), size(Peewee))”*[7]. SME can only match relations that have identical names. In this respect, HDTP shows more flexibility because relations can be matched irrespectively of having the same names or not.

One important feature of HDTP is that it is able to represent metaphors and analogies in domains that are more complex and cannot be represented in a simple algebraic way. Moreover, it provides more insight into the processes that underlie the establishment of the analogical relation than the theory presented in [25]. In Indurkhya’s approach, it is not examined further how exactly the cognitive relation is established. In HDTP, a general theory is computed, which is a description of both source and target. Such a generalization and the corresponding substitutions specify how the analogical relation is established. Another difference to the other approaches is that the analogical relation, induced by the algorithm HDTP-A, exists not only purely syntactically but also the semantic level is clearly specified.

In contrast to SME, HDTP also models the introduction of new concepts into the target domain. Thus, the processes of learning and inductive inference are modeled in a more appropriate way. In SME, this creativity is not taken into account because the concepts have to be given explicitly in the initial representations of the target domain.

In this section, two approaches to analogy were presented, which, as opposed to other approaches, also give formal specifications of the underlying structures of analogies. They have both been successfully applied to several examples of analogies and

[7]This holds for the implementation of SME that is presented in [8].

metaphors. This indicates that an algebraic framework seems to be appropriate to provide a formal base for models of analogy. In both approaches, the analogy is represented as a relation between two domains. These relations ensure a certain degree of structural similarity between source and target domain. In order to be able to characterize both relations, i.e. the cognitive relation and the analogical relation, more precisely, in the next section the concept of bisimulation will be presented. Bisimulation also describes structural similarity. Then the central relations of Indurkhya's theory and HDTP will be compared each other and to bisimulation.

4 Bisimulation

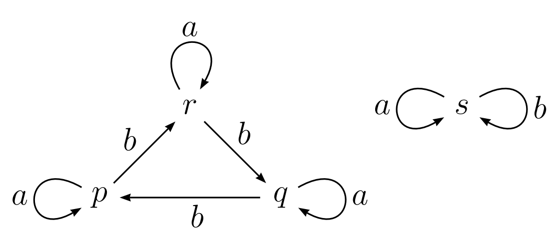
In this section, I will give an overview of the notion of bisimulation. Different domains will be presented in which bisimulations are used to express structural equivalences. A bisimulation is a binary relation that expresses structural identity without being isomorphic. The concept of bisimulation has its origins in the domain of concurrent processes and in modal logic; it is also used in various other fields, e.g. game theory and coalgebra.

4.1 Bisimulation in the Field of Concurrent Processes

The notion of bisimulation was first introduced by Park [28] and by Milner, who used it in the context of his calculus of communicating systems (CCS) [26, 27], a theoretical framework for modeling concurrent processes. In CCS, dynamic systems are modeled as *labeled transition systems.*

Definition 12 (Milner)**.** *A* ***labeled transition system*** $\mathcal{S}$ *is a triple* $\mathcal{S} = (S, A, \rightarrow_S)$*, where* S *is a set of states,* A *a set of labels (or actions) and* $\rightarrow_S \subseteq S \times A \times S$ *a set of transitions. Instead of writing* $\langle s, a, s' \rangle \in \rightarrow_S$*, usually the notation* $s \overset{a}{\longrightarrow}_S s'$ *is used.*

Labeled transition systems with a finite number of states can be represented in a nice way by directed graphs where the vertices represent the states of the system and the labeled arcs represent the transitions. Labeled transition systems can be nondeterministic: From one state there might be several transitions with the same label leading to different states.
Consider the following example:



When comparing both systems, it is quite obvious that their behavior is equivalent although they look different.
The motivation for introducing the notion of bisimulation is that for an external observer two systems seem to be equivalent if their behavior is identical.
This idea is represented in the notion of strong bisimilarity.

Definition 13 (Milner)**.** *Let $\mathcal{S} = (S, A, \rightarrow_S)$ and $\mathcal{T} = (T, A, \rightarrow_T)$ be labeled transition systems.*

1. *let $s_1 \in S$ and $t_1 \in T$, then $R \subseteq S \times T$ is a **strong bisimulation** for s_1 and t_1 if $\langle s_1, t_1 \rangle \in R$ and $\forall \langle s, t \rangle \in R, a \in A$ it holds that*
   (a) if $s \stackrel{a}{\longrightarrow}_S s'$, $\exists t' \in T$ with $t \stackrel{a}{\longrightarrow}_T t'$ and $\langle s', t' \rangle \in R$ and
   (b) if $t \stackrel{a}{\longrightarrow}_T t'$, $\exists s' \in S$ with $s \stackrel{a}{\longrightarrow}_S s'$ and $\langle s', t' \rangle \in R$.
2. *$\mathcal{S}$ and $\mathcal{T}$ are strongly bisimilar if there is a strong bisimulation for their initial states.*

In the example above, the relation $R = \{\langle p, s \rangle, \langle q, s \rangle, \langle r, s \rangle\}$ is a strong bisimulation between both systems.

Properties of Bisimulations [34]:

- Let $\mathcal{S} = (S, A, \rightarrow_S)$ be a labeled transition system. Then the identity relation id_S with $id_S(s) = s, \forall s \in S$, is a bisimulation for $\mathcal{S}$ and $\mathcal{S}$
- Let $\mathcal{T}_1, \mathcal{T}_2$ and $\mathcal{T}_3$ be labeled transition systems. If R is a strong bisimulation between $\mathcal{T}_1$ and $\mathcal{T}_2$ and R' is a strong bisimulation between $\mathcal{T}_2$ and $\mathcal{T}_3$, then the composition of R and R' is a strong bisimulation between $\mathcal{T}_1$ and $\mathcal{T}_3$.
- If R is a strong bisimulation for transition systems $\mathcal{T}_1$ and $\mathcal{T}_2$, R^{-1} is a strong bisimulation for $\mathcal{T}_2$ and $\mathcal{T}_1$.

Labeled transition systems can have a special transition that is referred to as *internal* or *silent transition*; it is symbolized as a transition having the label τ and represents actions of the system that are invisible to an external observer. By performing a silent transition, a system can change its current state without an observer noticing it.
In some cases, one is only interested in the observable behavior of a system. This is represented by writing $s \stackrel{a}{\Longrightarrow} s'$ for $s \stackrel{\tau^k}{\longrightarrow}\stackrel{a}{\longrightarrow}\stackrel{\tau^l}{\longrightarrow} s'$, with $k, l \in \mathbb{N}$, where $\stackrel{\tau^n}{\longrightarrow}$, $n \in \mathbb{N}$, stands for a sequence of n τ-transitions. Since silent transitions cannot be observed, it is reasonable to relax strong bisimilarity to a weaker notion of bisimilarity that is closer to what one would consider as *observational equivalence* [26] because it is only

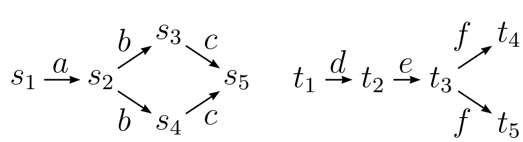
based on the observable behavior of a system. In cases where one has no information about the internal structure of a system, there is no other way of describing the system than describing it in terms of the observable behavior.

Definition 14 (Milner)**.** *Let $\mathcal{S} = (S, A, \rightarrow_S)$ and $\mathcal{T} = (T, A, \rightarrow_T)$ be labeled transition systems.*

1. *Let $s_1 \in S$ and $t_1 \in T$, then $R \subseteq S \times T$ is a* ***weak bisimulation*** *for s_1 and t_1 if $\langle s_1, t_1 \rangle \in R$ and $\forall \langle s, t \rangle \in R, a \in A$ it holds that*
   (a) If $s \overset{a}{\Longrightarrow}_S s'$, $\exists t' \in T$ with $t \overset{a}{\Longrightarrow}_T t'$ and $\langle s', t' \rangle \in R$ and
   (b) if $t \overset{a}{\Longrightarrow}_T t'$, $\exists s' \in S$ with $s \overset{a}{\Longrightarrow}_S s'$ and $\langle s', t' \rangle \in R$.
2. *$\mathcal{S}$ and $\mathcal{T}$ are weakly bisimilar if there is a weak bisimulation for their initial states.*

Weakly bisimilar systems are indistinguishable by an observer and can be seen as two black boxes that respond in the same way when one interacts with them[1] [2]. As investigated by Taubner in [34], there are various other notions of equivalences between states of transition systems, each of them capturing a different aspect of the systems. One example is *trace equivalence*: Two states are trace equivalent if they have the same possible sequences of actions[2] starting from them. It has been found out that apart from isomorphism, strong bisimilarity is the finest equivalence for states of a labeled transition system (cf. [34] for a proof). This shows that strong bisimilarity is a relatively strong notion for expressing equivalences. Depending on the kind of dynamic processes under consideration and the characteristics that are of interest, different variations of bisimulations exist. In [38], a hierarchy of 155 variants of bisimulations and bisimulation-like equivalences is developed.
Examples such as the following one illustrate that it is reasonable to consider a more general notion of bisimulation that allows the labels of the outgoing transitions of bisimilar states to be nonidentical.

Example 1. Let $\mathcal{S} = (S, A_S, \rightarrow_S)$ and $\mathcal{T} = (T, A_T, \rightarrow_T)$ be labeled transition systems represented by the following graphs:



[1] Obviously, the same holds for strongly bisimilar systems.
[2] A possible sequence of actions is a sequence of labels corresponding to a sequence of transitions that can be performed starting in this state.

Since there is no pair of identical labels, a bisimulation cannot be established, but nevertheless, the structures of both systems seem to be equivalent on a more general level. This idea is investigated more formally in [11], where the definition of bisimulation is relaxed and weak and strong *generalized bisimulation with respect to a relation* are introduced.

Definition 15 (Galpin). *Let $\mathcal{S} = (S, A_S, \rightarrow_S)$ and $\mathcal{T} = (T, A_T, \rightarrow_T)$ be labeled transition systems and let $B \subseteq A_S \times A_T$. $R \subseteq S \times T$ is a **strong generalized bisimulation with respect to B** such that $\langle s, t\rangle \in R$ only if*

1. *For all $a_S \in A_S$, whenever $s \xrightarrow{a_S}_S s'$, there exists $t' \in T$ and $a_T \in A_T$ such that $t \xrightarrow{a_T}_T t'$, $\langle a_S, a_T\rangle \in B$ and $\langle s', t'\rangle \in R$.*
2. *For all $a_T \in A_T$, whenever $t \xrightarrow{a_T}_T t'$, there exists $s' \in S$ and $a_S \in A_S$ such that $s \xrightarrow{a_S}_S s'$, $\langle a_S, a_T\rangle \in B$ and $\langle s', t'\rangle \in R$.*

In Example 1, $R = \{\langle s_1, t_1\rangle, \langle s_2, t_2\rangle, \langle s_3, t_3\rangle, \langle s_4, t_3\rangle, \langle s_5, t_4\rangle, \langle s_5, t_5\rangle\}$ is a strong generalized bisimulation with respect to $B = \{\langle a, d\rangle, \langle b, e\rangle, \langle c, f\rangle\}$.

In Section 5.1, it is shown that a variation of a generalized bisimulation with respect to a relation can be used to describe structural similarities of source concept network and environment in Indurkhya's framework.

4.2 Bisimulation from a Game Theoretic Perspective

Bisimulation is also a central notion in game theory. The relation between the notions of bisimulation we have considered so far and the one used in game theory is investigated by Stirling [32, 33]. Bisimulation in game theory is associated with *Ehrenfeucht-Fraïssé games*. These games are used in various applications in theoretical computer science [35].

One motivation for introducing a game characterization of bisimilarity is the following: Given a labeled transition system, how can we determine whether two states s and t are not bisimilar? According to the definition of bisimilarity of two states, one has to find all binary relations containing the pair $\langle s, t\rangle$ and check if they are bisimulations or not. If one can find a relation that is a bisimulation, s and t are bisimilar. Otherwise they are not. This method is very expensive: For a transition system with n states, one would have to check up to 2^{n^2} binary relations. As argued in [31], an alternative way to find out whether two states are not bisimilar is by playing a *bisimulation game*:

Definition 16 (Srba)**.** *Let $\mathcal{S} = (S, A, \rightarrow)$ be a labeled transition system. A* ***strong bisimulation game*** *$\mathcal{G}\langle s, t\rangle$ with initial configuration $\langle s, t\rangle$ with $s, t \in S$ is a sequence $\langle s, t\rangle, \langle s_1, t_1\rangle, \ldots$ with $s_i, t_i \in S$. Two players, Spoiler and Duplicator, play it. The game is played in rounds where each round consists of the following steps:*

1. *Spoiler makes the first move by choosing either*
   (a) *a transition $s \xrightarrow{\alpha} s_1$ or*
   (b) *a transition $t \xrightarrow{\alpha} t_1$.*
2. *Then it is Duplicator's turn. If Spoiler has chosen (a), he has to find a transition $t \xrightarrow{\alpha} t_1$, if Spoiler has chosen (b), Duplicator has to find a transition $s \xrightarrow{\alpha} s_1$.*
3. *$\langle s_1, t_1\rangle$ becomes the current configuration and the game continues as described above.*

Note the difference between the actions of the two players: Spoiler can choose one of the states of the current configuration and a transition starting from this state, whereas Duplicator can only choose a transition with the same label as Spoiler's transition. The game ends when one of the players cannot make a move; this player looses the game. Whenever the game is infinite, Duplicator wins. In other words: If Duplicator cannot find a move that matches Spoilers move, Spoiler wins. If Duplicator can match all of Spoilers moves or if the current configuration is a dead end (both states of the current configuration have no outgoing transitions), Duplicator wins. Of course, a bisimulation game can also be played on two different systems that have the same sets of labels; the game can be played on the disjoint union of both systems.

The outcome of a bisimulation game reveals information about the structural similarity of the states of the initial configuration. The game theoretic background of this is given in [35].

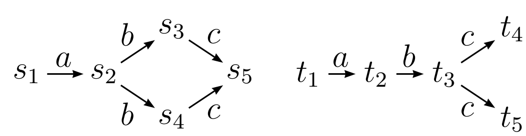
Definition 17 (Stirling)**.** *Let $\mathcal{S} = (S, A, \rightarrow_S)$ be a labeled transition system. $s, t \in S$ are said to be* ***strongly game equivalent*** *if and only if Duplicator has a universal winning strategy[3] for the game $\mathcal{G}\langle s, t\rangle$.*

An important result is that strong game equivalence is equivalent to strong bisimilarity: Two states of a labeled transition system are strongly game equivalent if and only if they are strongly bisimilar.
Let $\mathcal{S} = (S, A, \rightarrow)$ be a transition system and let $s, t \in S$. Assume s and t are strongly game equivalent. In order to show that there is a bisimulation between them, let $R = \{\langle s, t\rangle |\ s \textit{ and } t \textit{ are strongly game equivalent}\}$. Since s and t are

[3]Duplicator having a universal winning strategy means that he can always win the game irrespectively of the moves chosen by Spoiler.

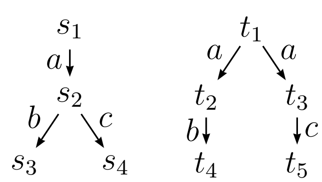
strongly game equivalent, we know that Duplicator can win all games $\mathcal{G}\langle s,t\rangle$. This means that for all outgoing transitions from either s or t there is a corresponding one from the other state such that the resulting states s' and t' are again strongly game equivalent and thus, $\langle s',t'\rangle \in R$. Hence, R is a bisimulation.
Now assume that R is a bisimulation and $\langle s,t\rangle \in R$. Then in the game $\mathcal{G}\langle s,t\rangle$ Duplicator can respond to all of Spoilers moves in a way that the resulting configuration is a pair of states that is in R as well because that is exactly what the definition of bisimulations says. Therefore, whenever Spoiler is able to make a move, Duplicator can respond with a corresponding one and the game continues until Spoiler is stuck. Hence, s and t are strongly game equivalent.

Example 2. In these graphs, s_1 and t_1 are strongly game equivalent.



It can easily be seen that Duplicator can win every game $\mathcal{G}\langle s_1,t_1\rangle$.

Example 3. Here, s_1 and t_1 are not game equivalent.



When the game $\mathcal{G}\langle s_1,t_1\rangle$ is in configuration $\langle s_2,t_2\rangle$, Spoiler can simply choose $s_2 \xrightarrow{c} s_4$ and Duplicator cannot match her move. In configuration $\langle s_2,t_3\rangle$, Spoiler can choose $s_2 \xrightarrow{b} s_3$ and again, Duplicator cannot respond. So, Spoiler can win all games $\mathcal{G}\langle s_1,t_1\rangle$ and s_1 and t_1 are not bisimilar.

These examples show that game theory provides a quite simple and illustrative way to check if two states of a transition system are strongly bisimilar or not[4].
We can also introduce a game characterization of a generalized strong bisimulation with respect to a relation.

Definition 18. *Let $\mathcal{S} = (S, A_S, \rightarrow_S)$ and $\mathcal{T} = (T, A_T, \rightarrow_T)$ be labeled transition systems and let $B \subseteq A_S \times A_T$ be a relation on the sets of labels. A* ***generalized strong bisimulation game with respect to the relation B*** *$\mathcal{G}_B\langle s,t\rangle$ with initial*

[4] A *weak bisimulation game* can be defined analogously by replacing the $\longrightarrow$ with $\Longrightarrow$ in Definition 16.

configuration $\langle s,t\rangle, s \in S, t \in T$ *is a sequence* $\langle s,t\rangle, \langle s_1,t_1\rangle, \ldots$ *with* $s_i \in S, t_i \in T$. *Two players, Spoiler and Duplicator, play it. It is played in rounds where each round consists of the following steps:*

1. *Spoiler makes the first move by choosing either*
   (a) *a transition* $s \xrightarrow{\alpha} s_1$ *or*
   (b) *a transition* $t \xrightarrow{\alpha} t_1$.
2. *Then it is Duplicator's turn. If Spoiler has chosen (a), Duplicator has to find a transition* $t \xrightarrow{\beta} t_1$ *with* $\langle \alpha, \beta\rangle \in B$.
   If Spoiler has chosen (b), Duplicator has to find a transition $s \xrightarrow{\beta} s_1$ *with* $\langle \beta, \alpha\rangle \in B$.
3. $\langle s_1, t_1\rangle$ *becomes the current configuration and the game continues as described above.*

Analogously to Definition 17, generalized strong game equivalence with respect to a relation can be defined which is, as one might expect, equivalent to strong generalized bisimilarity with respect to a relation. As an example, one can play a bisimulation game with respect to the relation B on the graphs given in Example 1.

4.3 Bisimulation from a Coalgebraic Perspective

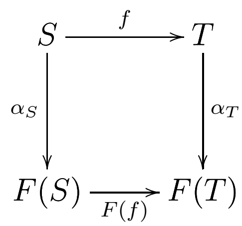
In this part, a more abstract approach to bisimulation via coalgebras is presented. Coalgebras are the dual of algebras. Whereas in algebra, one is concerned with how something is constructed, in coalgebra the interest is in decomposition and observable behavior. In contrast to algebras, we do not know how exactly a coalgebra is made up and it is more like a black box. One can just observe what happens when it breaks into parts. This is an important feature that shows that certain things can be modeled with coalgebras that cannot be modeled with algebras because of a lack of information about the exact construction of the entity under consideration.

As seen in the previous subsections, bisimulations are mainly used to describe equivalences based on the behavior of dynamic systems. State-based dynamic systems can be represented in a uniform way [18, 29]. All types of state-based systems are similar in that they consist of a set of states and a set of transitions that map sets of states to another set of states or to a set of states and an action, e.g. an output or a label. This can be formalized by the following map: $\alpha : S \to F(S)$, which is called *F-transition structure*. F is a *functor* that maps every set X to a new set $F(X)$ and that maps a map $f : X \to Y$, with X and Y being sets, to a new map $F(f) : F(X) \to F(Y)$.

Definition 19 (Gumm)**.** *A **coalgebra of type F** is a pair $\langle S, \alpha_S \rangle$ with S being a set and α_S a map with $\alpha_S : S \to F(S)$.*

For a labeled transition system $\mathcal{S} = (S, \longrightarrow_S, A)$ we have the functor $F_L(S) \subseteq A \times S$. Then a labeled transition system $(S, \longrightarrow_S, A)$ can be represented as a tuple $\langle S, \alpha_S \rangle$ with $\alpha_S : S \to F_L(S)$, $s \mapsto \{\langle a, s' \rangle | s \xrightarrow{a}_S s'\}$.
Now a homomorphism between systems (coalgebras) can be defined:

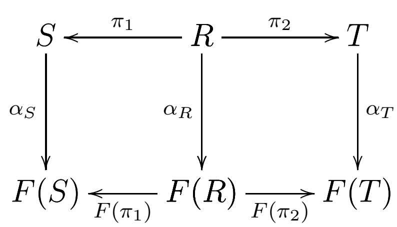
Definition 20 (Rutten)**.** *A **homomorphism** between coalgebras of type F, $\langle S, \alpha_S \rangle$ and $\langle T, \alpha_T \rangle$, is a function $f : S \to T$ such that $\alpha_T \circ f = F(f) \circ \alpha_S$, i.e. the following diagram commutes:*



In Subsection 4.1, bisimulation was defined as a relation between states of a system that are indistinguishable by an external observer. For coalgebras, bisimulations are defined in the following way:

Definition 21 (Rutten)**.** *A **bisimulation** between coalgebras of type F, $\langle S, \alpha_S \rangle$ and $\langle T, \alpha_T \rangle$, is a binary relation $R \subseteq S \times T$ such that a transition structure $\alpha_R : R \to F(R)$ can be defined, with projections $\pi_1 : R \to S$, $\langle s, t \rangle \mapsto s$ and $\pi_2 : R \to T$, $\langle s, t \rangle \mapsto t$ being homomorphisms.*

This can be illustrated as a commuting diagram:

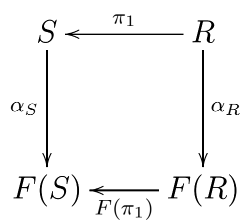


This coalgebraic version of bisimulation seems less intuitive than the ones given in the previous subsections. Comparing a bisimulation for labeled transitions systems to the coalgebraic one shows that both definitions of bisimulation are equivalent:
Let $\mathcal{S} = (S, \longrightarrow_S, A)$ and $\mathcal{T} = (T, \longrightarrow_T, A)$ be labeled transition systems and let $R \subseteq S \times T$ be a bisimulation between $\mathcal{S}$ and $\mathcal{T}$. Then it holds that $\forall \langle s, t \rangle \in R :$

1. For all $s' \in S$, if $s \xrightarrow{a}_S s'$, there is a $t' \in T$ with $t \xrightarrow{a}_T t'$ and $\langle s', t' \rangle \in R$.

2. For all $t' \in T$, if $t \xrightarrow{a}_T t'$, there is a $s' \in S$ with $s \xrightarrow{a}_S s'$ and $\langle s', t' \rangle \in R$.

The systems can be represented as coalgebras $\langle S, \alpha_S \rangle$ and $\langle T, \alpha_T \rangle$ with
$\alpha_S : S \to F(S)$, $s \mapsto \{\langle a, s' \rangle | s \xrightarrow{a}_S s'\}$ and $\alpha_T : T \to F(T)$, $t \mapsto \{\langle a, t' \rangle | t \xrightarrow{a}_T t'\}$.
Now let α_R be a transition structure with
$\alpha_R : R \to F(R)$, $\langle s, t \rangle \mapsto \{\langle s', t' \rangle | s \xrightarrow{a}_S s'$ and $t \xrightarrow{a}_T t'\}$.
Then, $\forall \langle s, t \rangle \in R$ it holds that $\alpha_S(\pi_1 \langle s, t \rangle) = \alpha_S(s) = \{\langle a, s' \rangle | s \xrightarrow{a}_S s'\}$ and
$F(\pi_1)(\alpha_R \langle s, t \rangle) = F(\pi_1)(\{\langle s', t' \rangle | s \xrightarrow{a}_S s'$ and $t \xrightarrow{a}_T t'\}) = \{\langle a, s' \rangle |\ s \xrightarrow{a}_S s'\}$.
Thus, $\alpha_S \circ \pi_1 = F(\pi_1) \circ \alpha_R$ and π_1 is a homomorphism. Analogously, it can be shown that π_2 is a homomorphism.
For the converse direction, let $R \subseteq S \times T$ with transition structure $\alpha_R : R \to F(R)$ fulfill the coalgebraic conditions for a bisimulation between $\langle S, \alpha_S \rangle$ and $\langle T, \alpha_T \rangle$. Are the conditions 1 and 2 for a bisimulation between transition systems satisfied?
Let $\langle s, t \rangle \in R$ and $s \xrightarrow{a}_S s'$.
Then $s = \pi_1 \langle s, t \rangle$ and $\pi_1 \langle s, t \rangle \xrightarrow{a}_S s'$. Since π_1 is a homomorphism, the following diagram commutes:



and there is a pair $\langle s'', t' \rangle \in R$ with $\langle s, t \rangle \xrightarrow{a}_R \langle s'', t' \rangle$ and $\pi_1 \langle s'', t' \rangle = s'$. Thus, $s'' = s'$ and $\langle s', t' \rangle \in R$.
It remains to show that $t \xrightarrow{a}_T t'$. Since we already know that $\langle s', t' \rangle \in R$ and that π_2 is a homomorphism and $\alpha_T(\pi_2 \langle s, t \rangle) = F(\pi_2)(\alpha_R \langle s, t \rangle) = \{\langle a, t' \rangle | t \xrightarrow{a}_T t'\}$, it holds that $t \xrightarrow{a}_T t'$. Hence, condition 1 for a bisimulation is satisfied. Analogously, it can be shown that condition 2 is satisfied as well.

4.4 Bisimulation in Modal Logic

Independently from the notions of bisimulation described in the previous subsections, already in 1976 bisimulation was introduced in modal logic by van Benthem as *p-relation* [36]. Note that the primary concern in this subsection is to get an understanding of the notion of bisimulation in modal logic, and therefore I will only give a short introduction to modal logic that covers only the concepts that are necessary to get an idea of bisimulation in this field. In [2], bisimulations in modal logic are studied in more detail and are used to establish various results.

Whereas classical first-order logic discriminates propositions according to whether they are true or false, one motivation for introducing modal logic was that in natural

language we do not only distinguish between 'true' and 'false' but are able to make finer distinctions. Propositions also differ in how they are true or false. Instead of adopting a global view, in modal logic functions are evaluated locally inside a structure.
One important concept is that of a *relational structure*. A relational structure is a nonempty set together with a relation defined on it. If we think about labeled transition systems as sets of states with $\xrightarrow{a}$ being a binary relation on this set, they can be seen as simple relational structures as well.
For our purpose it is sufficient to consider only the basic modal language and not extensions of it.

Definition 22 (Blackburn et al.)**.** *The* ***basic modal language*** *is defined using* Φ*, which stands for a set of proposition letters, and a unary modal operator* $\Diamond$ *('diamond'). The well-formed formulas* ϕ *of the basic modal language are given by the following rule:*

$$\phi := p \mid \bot \mid \neg\phi \mid \psi \vee \phi \mid \Diamond\phi,$$

where p *is an element of* Φ*,* $\bot$ *represents the constant falsum and* ψ *is a well-formed formula.*

The dual of $\Diamond$ is $\Box$ ('box'), with $\Box\phi := \neg\Diamond\neg\phi$.
There are different readings and interpretations of the modal operators; one of the most influential ones is reading $\Diamond\phi$ as 'it is possibly the case that ϕ' and $\Box\phi$ as 'it is necessarily the case that ϕ'. The branch of modal logic where the modal operators are interpreted this way is called *alethic* modal logic. As mentioned above, in modal logic functions are evaluated inside relational structures. The following definitions make the relational structures in which the basic modal language is interpreted, more precise:

Definition 23 (Blackburn et al.)**.** *A* ***frame*** *for the basic modal language is a pair* $\mathfrak{F} = (W, R)$ *with* $W \neq \emptyset$ *and* $R \subseteq W \times W$*.*

W is often referred to as 'universe' and the elements of W are called 'worlds' or 'states'. A *model* extends a frame by contingent information about the elements of the universe W.

Definition 24 (Blackburn et al.)**.** *A* ***model*** *for the basic modal language is a pair* $\mathfrak{M} = (\mathfrak{F}, V)$*, where* $\mathfrak{F}$ *is a frame (W,R) and* V *is a valuation function* $V : \Phi \to \mathcal{P}(W)$*. On an informal level, one can say that* V *maps a proposition* p *to the set of states where* p *is true.*

Models and frames are both special types of relational structures. In order to see how truth and falsity of formulas within states of a model are defined, consider the following definition:

Definition 25 (Blackburn et al.). *Let w be a state in a model $\mathfrak{M} = (W, R, V)$ and let ϕ be a formula. Satisfaction of ϕ is defined inductively:*

$\mathfrak{M}, w \models p$	*if and only if*	$w \in V(p)$
$\mathfrak{M}, w \models \bot$		*never*
$\mathfrak{M}, w \models \neg\phi$	*if and only if*	*not* $\mathfrak{M}, w \models \phi$
$\mathfrak{M}, w \models \phi \vee \psi$	*if and only if*	$\mathfrak{M}, w \models \phi$ *or* $\mathfrak{M}, w \models \psi$
$\mathfrak{M}, w \models \Diamond\phi$	*if and only if*	*for some* $v \in W$ *with* wRv *it holds that* $\mathfrak{M}, v \models \phi$

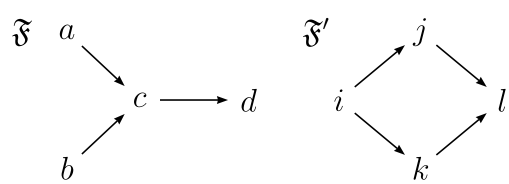
It follows that $\mathfrak{M}, w \models \Box\phi$ if and only if for all v with wRv it holds that $\mathfrak{M}, v \models \phi$.

Models are not primarily studied in isolation, but what one is usually interested in are relations between models and the properties that are invariant under such relations. It is this context where bisimulations play a central role in modal logic. Roughly speaking, a bisimulation is a relation between two models such that related states carry identical atomic information and have equivalent transition possibilities [2].

Definition 26 (Blackburn et al.). *Let $\mathfrak{M} = (W, R, V)$ and $\mathfrak{M}' = (W', R', V')$ be models. $Z \subseteq W \times W', Z \neq \emptyset$ is called a **bisimulation** between $\mathfrak{M}$ and $\mathfrak{M}'$ if the following conditions are satisfied:*

1. *If wZw', w and w' satisfy the same propositional letters.*
2. *If wZw' and wRv, there exists $v' \in W'$ such that $w'Rv'$ and vZv'.*
3. *If wZw' and $w'Rv'$, there exists $v \in W$ such that wRv and vZv'.*

Example 4. Let $\mathfrak{M} = (\mathfrak{F}, V)$ *and* $\mathfrak{M}' = (\mathfrak{F}', V')$ be models. The frames are represented as graphs:



$V(p) = \{a, b, d\}, V(q) = \{c\}, V'(p)\{i, l\}, V'(q) = \{j, k\}$.
Then $Z = \{\langle a, i\rangle, \langle b, i\rangle, \langle c, j\rangle, \langle c, k\rangle, \langle d, l\rangle\}$ is a bisimulation between $\mathfrak{M}$ and $\mathfrak{M}'$.

In [17], the relation between modal logic, concurrent processes and coalgebras is studied in more detail and on a more formal level. As seen in Section 4.1, bisimulation describes equivalences of state-based systems. Hennessy and Milner [21] applied modal logic to process algebra and used it to reason about state-based systems. They developed a modal language for this purpose. In this language, two states are bisimilar, i.e. observationally equivalent, if and only if they satisfy the same modal formulas.

4.5 Summarizing Bisimulations

This section has shown that bisimulation is a powerful notion for describing equivalences of state-based systems. Such systems can be represented as coalgebras, and structural equivalences of coalgebras can be described by bisimulations. Another abstract way of representing state-based systems is as relational structures. Modal logic provides various tools for reasoning about relational structures. In modal logic, bisimulation is not only used as a structure preserving relation between models but is a key notion which plays a central role in important theorems. In its game theoretic version, bisimilarity is characterized by the outcome of Ehrenfeucht-Fraïssé games that can be played on graph-like structures.

There is another field where bisimulations are used, which is not mentioned in this section. This is *non-well-founded set theory* [1], which allows sets to be members of themselves. Such sets can be represented as graphs and are said to be bisimilar if their graphs are bisimilar.

The character of bisimilarity as structural equivalence and the fact that it is successfully used in various domains justify that it could be reasonable to investigate how far analogies can be seen as bisimulations since, as explained in Section 2, analogy is a way to establish and describe structural similarities, too.

5 Analogy as Bisimulation

In this section, the idea that bisimulation might be suitable for describing analogies will be investigated on a formal level by comparing the central relations, modeling the connection between source and target, of the approaches presented in Sectionn 3 to bisimulation. In Indurkhya's theory, a coherent cognitive relation ensures structural similarity of source and target. The counterpart in HDTP is the analogical relation, which is induced by the algorithm HDTP-A.

5.1 Coherency and Bisimilarity

Coherent cognitive relations and the notions of bisimulation we have considered so far are from different domains. For this reason, we cannot compare them directly but will first consider how a bisimulation in the domain of Indurkhya's cognitive models would look like.

5.1.1 Bisimulation in Indurkhya's Framework

The notion of bisimulation has to be translated into the language of Indurkhya's theory. One difference between a coherent cognitive relation and the classical definition of bisimulation is that a bisimulation is a binary relation over sets of states, whereas the cognitive relation in Indurkhya's cognitive models relates operators as well. In Indurkhya's framework, there are no transitions but n-ary operators; performing a transition starting from a certain state in a labeled transition system corresponds to applying an n-ary operator to an n-tuple of objects of the source concept network or the environment algebra. The idea of bisimilarity of two states of a transition system is that for all transitions from one state there is a corresponding one from the other state such that the results of these transitions are bisimilar states as well. Here the term 'corresponding' refers to transitions having equivalent labels.
What could be the respective condition for cognitive models? Given a pair of bisimilar n-tuples of objects and applying an operator to one of them, applying a corresponding one to the other n-tuple means here applying an operator that is related to the first

one (via Ψ). Formalizing these ideas results in the following definition[1]:

Definition 27. *Given a cognitive model* $\mathcal{C} = \langle\langle A, \Omega\rangle, \langle R, \Psi\rangle, \langle B, \Sigma\rangle\rangle$*, we say that* $\langle R, \Psi\rangle$ *is* ***a bisimulation*** *between* $\langle A, \Omega\rangle$ *and* $\langle B, \Sigma\rangle$ *if and only if* $\forall$ *arities n:* $\forall x_1, \ldots, x_n \in A,\ y_1, \ldots, y_n \in B$ *it holds that whenever* $\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle \in R$*, the following conditions are satisfied:*

1. $\forall \omega \in \Omega(n)$ *it is the case that whenever* $\omega(x_1, \ldots, x_n) \in A$*,* $\exists \sigma \in \Sigma(n)$ *such that* $\langle \omega, \sigma\rangle \in \Psi(n)$ *and* $\langle \omega(x_1, \ldots, x_n), \sigma(y_1, \ldots, y_n)\rangle \in R$*.*

2. $\forall \sigma \in \Sigma(n)$ *it is the case that whenever* $\sigma(y_1, \ldots, y_n) \in B$*,* $\exists \omega \in \Omega(n)$ *such that* $\langle \omega, \sigma\rangle \in \Psi(n)$ *and* $\langle \omega(x_1, \ldots, x_n), \sigma(y_1, \ldots, y_n)\rangle \in R$*.*

5.1.1.1 An Example

This example will illustrate the formal concept of bisimulation in Indurkhya's framework. We will use a simplified version of an example presented in [25], where Indurkhya applies his framework to the cognitive processes of an agent called 'Spinner' that lives in a two-dimensional world *Flatland* (Figure 5.1), which is similar to the world in Edwin Abbott's novel *Flatland*. Flatland is divided into sections of equal width and all Flatland is pervaded by uniform diffuse light. The objects of the world surrounding Spinner are straight lines of a fixed length equal to the width of the sections. In each section there is one line. These lines can spin around their center points. Whenever they are in rest, they are in one of the following positions: —,/, \ or |. Spinner perceives its environment via a light sensitive sensory organ. This 'eye' is of the same width as the sections. Whenever there is a line in front of Spinner's eye, it causes a shadow on the eye, and thereby Spinner perceives the line in its visual field. Because of the difference in size of the shadows, Spinner is able to distinguish between lines in the positions —, | and the diagonal positions. The sensory information Spinner gets when lines in the orientations / or \ are in its visual field is identical because lines in these positions cause shadows of equal width. For this reason, Spinner is not able to distinguish between lines in the orientations / and \. Nevertheless, Spinner has conceptualizations of the different orientations / and \. Spinner's eye also serves as an effectory organ and thereby allows for interaction with the environment (i.e. with the lines). Spinner can emit little jet streams of air via this effectory organ and can thereby cause the lines to spin. The intensity of the jet streams can be varied in two levels. The jet streams run in parallel to the borders of the sections of Flatland. They are emitted either left or right from the center of Spinner's eye. Although Spinner knows that the jet streams are emitted either left or

[1]In Indurkhya's framework there is nothing like a silent transition and therefore we do not need to distinguish between strong and weak bisimilarity.

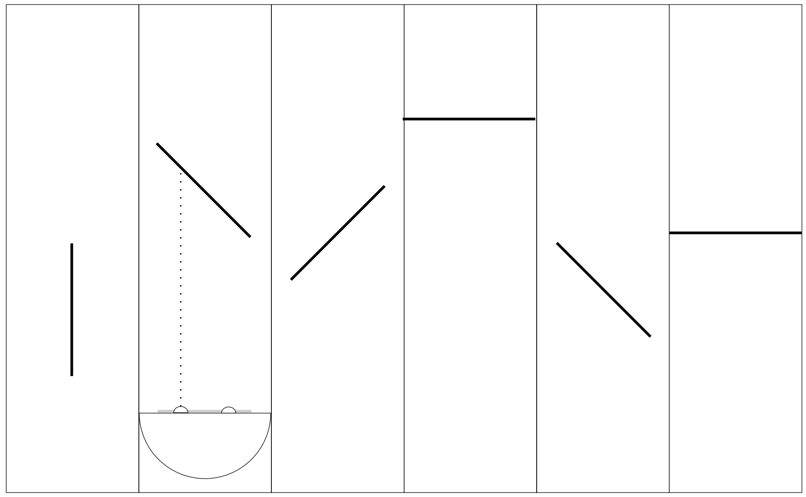


Figure 5.1: Spinner in Flatland

right of the center of its eye, it is not able to perceive or control the side of emittance, which is totally arbitrary. Spinner is also able to move through Flatland by jumping from one section to an adjacent one.

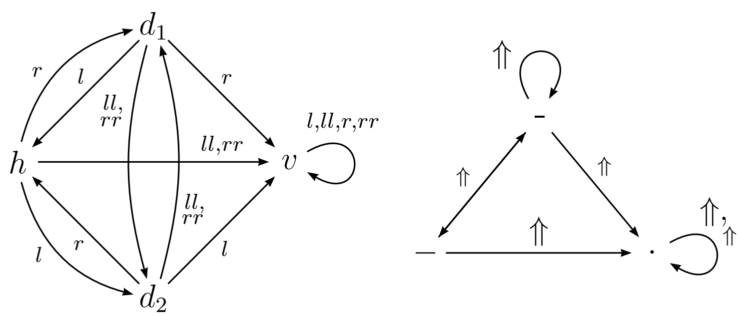
When formalizing Spinner's conceptualization of the environment, one gets four different symbols, representing the possible orientations of the lines. Let h, d_1, d_2 and v represent the orientations —, /, \ and | respectively.
As sensory information let — represent the shadow of horizontal line, $-$ that of a diagonal line and let $\cdot$ denote the sensory input Spinner gets when a vertical line causes a shadow on its eye.
Let us now formalize the air puffs that cause the spinning of the lines. As mentioned above, Spinner has a conceptualization of the two different intensities of the air puffs and of the side where they are emitted. An air puff of the lower intensity level emitted from the right side of the eye is represented by r and one emitted from the left side of the eye by l. The ones of the higher intensity level are denoted by rr and ll respectively.
Since Spinner is only able to control and distinguish the intensities of the air puffs, let the representation of the motor information consist of $\scriptstyle\Uparrow$ for the lower intensity level and $\Uparrow$ for the higher one.
Combining the representations results in the following source concept network:
$\mathscr{A} = \langle A, \Omega \rangle$,
$A = \{h, d_1, d_2, v\}$, $\Omega = \{r, l, rr, ll\}$, where all operators are of arity 1.

The environment is represented as $\mathscr{B} = \langle B, \Sigma \rangle$,

$B = \{—, -, \cdot\}$, $\Sigma = \{{\scriptstyle\Uparrow}, \Uparrow\}$, with ${\scriptstyle\Uparrow}$ and $\Uparrow$ being of arity 1.

The connection between Spinner's conceptualization of the world and its sensorimotor information can be formalized as a cognitive relation: $\mathscr{R} = \langle R, \Psi\rangle$,
$R = \{\langle h, —\rangle, \langle d_1, -\rangle, \langle d_2, -\rangle, \langle v, \cdot\rangle\}$, $\Psi = \{\langle r, {\scriptstyle\Uparrow}\rangle, \langle l, {\scriptstyle\Uparrow}\rangle, \langle rr, \Uparrow\rangle, \langle ll, \Uparrow\rangle\}$.

Spinner's air puffs have different effects on the lines, depending on their intensity, their side of emittance and the orientation of the lines. When a line is in the vertical position $|$, obviously there is no way for Spinner to make it spin because the air will either pass the line on the left or on the right. In all other orientations, let an air puff of low intensity make a line spin around its center point 45° in the respective direction depending on the side of emittance of the air puff. An air blast of high intensity causes the line to spin 90°. A representation of source concept network and environment as directed graphs will illustrate this:



Spinner's cognitive relation satisfies the conditions for a bisimulation as defined in Definition 27. All operators in Spinner's cognitive model are of arity 1, and in such cases a bisimulation in Indurkhya's theory is a generalized bisimulation with respect to the relation between the operators (cf. Definition 15). In this example, R is a generalized bisimulation with respect to Ψ. Duplicator can win all games $\mathcal{G}_\Psi\langle a, b\rangle$ with $\langle a, b\rangle \in R$. He can match all of Spoiler's moves and the game continues. Once it reaches the configuration $\langle v, \cdot\rangle$, there is no way to leave this configuration, and the game goes on forever.

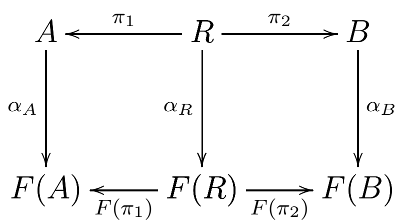
If Spinner's cognitive relation is a bisimulation, it must also satisfy the conditions of the coalgebraic version of bisimulation. $\mathscr{R}$ has to be formulated as a transition structure such that the projections from $R \subseteq A \times B$ are homomorphisms.
Source concept network and environment can be represented as transition structures:
$\alpha_A : A \to F(A)$, $x \mapsto \{\langle \omega, x'\rangle | x \xrightarrow{\omega} x'\}$,
$h \mapsto \{\langle l, d_2\rangle, \langle r, d_1\rangle, \langle ll, v\rangle, \langle rr, v\rangle\}$,
$d_1 \mapsto \{\langle l, h\rangle, \langle r, v\rangle, \langle ll, d_2\rangle, \langle rr, d_2\rangle\}$,

$d_2 \mapsto \{\langle l, v\rangle, \langle r, h\rangle, \langle ll, d_1\rangle, \langle rr, d_1\rangle\}$,
$v \mapsto \{\langle l, v\rangle, \langle r, v\rangle, \langle ll, v\rangle, \langle rr, v\rangle\}$,

$\alpha_B : B \rightarrow F(B)$, $y \mapsto \{\langle \sigma, y'\rangle | y \xrightarrow{\sigma} y'\}$,
$\text{---} \mapsto \{\langle _{\Uparrow}, -\rangle, \langle \Uparrow, \cdot\rangle\}$,
$- \mapsto \{\langle _{\Uparrow}, \text{---}\rangle, \langle _{\Uparrow}, \cdot\rangle, \langle \Uparrow - \rangle\}$,
$\cdot \mapsto \{\langle _{\Uparrow}, \cdot\rangle, \langle \Uparrow, \cdot\rangle\}$.

The cognitive relation can also be formalized as a transition structure:
$\alpha_R : R \rightarrow F(R)$, with $R \subseteq A \times B, \Psi \subseteq \Omega \times \Sigma$ and $\langle x, y\rangle \mapsto \{\langle\langle \omega, \sigma\rangle \langle x', y'\rangle\rangle | x \xrightarrow{\omega} x'$ and $y \xrightarrow{\sigma} y'\}$,
$\langle h, \text{---}\rangle \mapsto \{\langle\langle l, _{\Uparrow}\rangle, \langle d_2, -\rangle\rangle, \langle\langle r, _{\Uparrow}\rangle, \langle d_1, -\rangle\rangle, \langle\langle ll, \Uparrow\rangle, \langle v, \cdot\rangle\rangle, \langle\langle rr, \Uparrow\rangle, \langle v, \cdot\rangle\rangle\}$,
$\langle d_1, -\rangle \mapsto \{\langle\langle l, _{\Uparrow}\rangle, \langle h, \text{---}\rangle\rangle, \langle\langle r, _{\Uparrow}\rangle, \langle v, \cdot\rangle\rangle, \langle\langle ll, \Uparrow\rangle, \langle d_2, -\rangle\rangle, \langle\langle rr, \Uparrow\rangle, \langle d_2, -\rangle\rangle\}$,
$\langle d_2, -\rangle \mapsto \{\langle\langle l, _{\Uparrow}\rangle, \langle v, \cdot\rangle\rangle, \langle\langle r, _{\Uparrow}\rangle, \langle h, \text{---}\rangle\rangle, \langle\langle ll, \Uparrow\rangle, \langle d_1, -\rangle\rangle, \langle\langle rr, \Uparrow\rangle, \langle d_1, -\rangle\rangle\}$,
$\langle v, \cdot\rangle \mapsto \{\langle\langle l, _{\Uparrow}\rangle, \langle v, \cdot\rangle\rangle, \langle\langle r, _{\Uparrow}\rangle, \langle v, \cdot\rangle\rangle, \langle\langle ll, \Uparrow\rangle, \langle v, \cdot\rangle\rangle, \langle\langle rr, \Uparrow\rangle, \langle v, \cdot\rangle\rangle\}$.

It is the case that $\forall \langle x, y\rangle \in R : \alpha_A(\pi_1\langle x, y\rangle) = F(\pi_1)(\alpha_R\langle x, y\rangle)$ *and* $\alpha_B(\pi_2\langle x, y\rangle) = F(\pi_2)(\alpha_R\langle x, y\rangle)$.
Thus, π_1 and π_2 are homomorphisms and the following diagram commutes:



Hence, R is a bisimulation between $\langle A, \alpha_A\rangle$ and $\langle B, \alpha_B\rangle$.

The previous example illustrates that our notion of bisimulation in Indurkhya's framework seems to be appropriate.

5.1.2 Is a Coherent Cognitive Relation a Bisimulation?

Comparing a coherent cognitive relation to the bisimulation given in Definition 27 results in the following fact:

Proposition 1. *Given a coherent cognitive model* $\mathcal{C} = \langle\langle A, \Omega\rangle, \langle R, \Psi\rangle, \langle B, \Sigma\rangle\rangle$, *the cognitive relation* $\langle R, \Psi\rangle$ *is a bisimulation if for all arities* n *it holds that* $\Psi(\omega) \neq \emptyset, \forall \omega \in \Omega(n)$.

Proof. Let $\mathcal{C} = \langle\langle A, \Omega\rangle, \langle R, \Psi\rangle, \langle B, \Sigma\rangle\rangle$ be a coherent cognitive model with $\Psi(\omega) \neq \emptyset, \forall \omega \in \Omega(n)$ and let $\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle \in R$.

Let $\omega \in \Omega(n)$ with $\omega(x_1, \ldots, x_n) \in A$. Then it holds that $\exists \sigma \in \Sigma(n)$ with $\langle \omega, \sigma \rangle \in \Psi(n)$, and since $\langle R, \Psi \rangle$ is coherent, it is the case that $\langle \omega(x_1, \ldots, x_n), \sigma(y_1, \ldots, y_n) \rangle \in R$ and the first condition for a bisimulation is satisfied.
Now let $\sigma' \in \Sigma(n)$ with $\sigma'(y_1, \ldots, y_n) \in B$. Because of $\langle R, \Psi \rangle$ being a cognitive relation, $\exists \omega' \in \Omega(n)$ with $\langle \omega', \sigma' \rangle \in \Psi(n)$ (cf. Definition 2), and since $\langle R, \Psi \rangle$ is coherent, it holds that $\langle \omega'(x_1, \ldots, x_n), \sigma'(y_1, \ldots, y_n) \rangle \in R$ and condition 2 for a bisimulation is satisfied as well. Hence, $\langle R, \Psi \rangle$ is a bisimulation. □

This shows that a coherent cognitive relation is a bisimulation if on the operator level it is defined on the whole source.
Remember from Section 3.1 that one of the properties of a cognitive model is that the relevant subset of the source concept network is a finitely generated concept network.

Corollary 1. *Given a coherent cognitive model* $\mathcal{C} = \langle \langle A, \Omega \rangle, \langle R, \Psi \rangle, \langle B, \Sigma \rangle \rangle$, *let* $\langle A', \Omega' \rangle$ *denote the subset of the source concept network that is relevant, i.e.* $\langle A', \Omega' \rangle = \langle R, \Psi \rangle^{-1}(\langle B, \Sigma \rangle)$. *Then* $\mathcal{C}' = \langle \langle A', \Omega' \rangle, \langle R, \Psi \rangle, \langle B, \Sigma \rangle \rangle$ *is also a coherent cognitive model and* $\langle R, \Psi \rangle$ *is a bisimulation between* $\langle A', \Omega' \rangle$ *and* $\langle B, \Sigma \rangle$.

It is clear that $\mathcal{C}'$ is a coherent cognitive model if $\mathcal{C}$ is coherent because the coherency condition constrains only the relevant elements of the source concept network. It follows from Proposition 1 that $\langle R, \Psi \rangle$ is a bisimulation between $\langle A', \Omega' \rangle$ and $\langle B, \Sigma \rangle$ because obviously for all arities n it holds that $\Psi(\omega') \neq \emptyset, \forall \omega' \in \Omega'(n)$.

To summarize the results of the comparison of a coherent cognitive relation and bisimulation so far, we can say that a coherent cognitive relation is a bisimulation between the relevant part of the source concept network and the environment.

5.1.3 Is a Bisimulation Coherent?

The next step is to consider the converse direction, i.e. is a bisimulation between source concept network and environment coherent? Let $\mathcal{C} = \langle \langle A, \Omega \rangle, \langle R, \Psi \rangle, \langle B, \Sigma \rangle \rangle$ be a cognitive model with $\langle R, \Psi \rangle$ being a bisimulation between $\langle A, \Omega \rangle$ and $\langle B, \Sigma \rangle$. Let $x_1, \ldots, x_n \in A, \omega \in \Omega(n)$ with $\omega(x_1, \ldots x_n) \in A$ and let $y_1, \ldots, y_n \in B$ such that $\langle x_1, y_1 \rangle, \ldots, \langle x_n, y_n \rangle \in R$ and let $\sigma \in \Sigma(n)$ with $\langle \omega, \sigma \rangle \in \Psi(n)$.
For $\langle R, \Psi \rangle$ being coherent it must be the case that $\langle \omega(x_1, \ldots, x_n), \sigma(y_1, \ldots, y_n) \rangle \in R$. Since $\langle R, \Psi \rangle$ is a bisimulation, there is at least one $\sigma' \in \Sigma(n)$ with $\langle \omega, \sigma' \rangle \in \Psi(n)$ and $\langle \omega(x_1, \ldots, x_n), \sigma'(y_1, \ldots, y_n) \rangle \in R$.
That is not enough for being coherent because coherency means that for *any* pair of related operators and related n-tuples of objects the results of applying the operators to the respective n-tuples are in R as well.
However, it is possible to specify under what additional conditions a bisimulation is coherent:

Proposition 2. *Given a cognitive model* $\mathcal{C} = \langle\langle A, \Omega\rangle, \langle R, \Psi\rangle, \langle B, \Sigma\rangle\rangle$ *with* $\langle R, \Psi\rangle$ *being a bisimulation between* $\langle A, \Omega\rangle$ *and* $\langle B, \Sigma\rangle$*. If* Ψ *or* Ψ^{-1} *are partial functions,* $\langle R, \Psi\rangle$ *is coherent.*

Proof. Let $\mathcal{C} = \langle\langle A, \Omega\rangle, \langle R, \Psi\rangle, \langle B, \Sigma\rangle\rangle$ be a cognitive model with $\langle R, \Psi\rangle$ being a bisimulation between $\langle A, \Omega\rangle$ and $\langle B, \Sigma\rangle$.

1. Let Ψ be a partial function.
   Since $\langle R, \Psi\rangle$ is a bisimulation, it holds that given $\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle \in R$ and $\omega \in \Omega(n)$ with $\omega(x_1, \ldots, x_n) \in A$, $\exists \sigma' \in \Sigma(n)$ with $\langle \omega, \sigma'\rangle \in \Psi(n)$ and $\langle \omega(x_1, \ldots, x_n), \sigma'(y_1, \ldots, y_n)\rangle \in R$. Since Ψ is a partial function, for each $\omega' \in \Omega(n)$ there is at most one $\sigma \in \Sigma(n)$ with $\langle \omega', \sigma\rangle \in \Psi(n)$. Therefore, $\forall \sigma \in \Sigma(n)$ with $\langle \omega, \sigma\rangle \in \Psi(n)$ it is the case that $\langle \omega(x_1, \ldots, x_n), \sigma(y_1, \ldots, y_n)\rangle \in R$, and $\langle R, \Psi\rangle$ is locally coherent in A.

2. Let Ψ^{-1} be a partial function.
   Since $\langle R, \Psi\rangle$ is a bisimulation, it holds that given $\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle \in R$ and $\sigma \in \Sigma(n)$ with $\sigma(y_1, \ldots, y_n) \in B$, $\exists \omega' \in \Omega(n)$ with $\langle \omega', \sigma\rangle \in \Psi(n)$ and $\langle \omega'(x_1, \ldots, x_n), \sigma(y_1, \ldots, y_n)\rangle \in R$. Since Ψ^{-1} is a partial function, for each $\sigma' \in \Sigma(n)$ there is at most one $\omega \in \Omega(n)$ with $\langle \omega, \sigma'\rangle \in \Psi(n)$. Thus, $\forall \omega \in \Omega(n)$ with $\langle \omega, \sigma\rangle \in \Psi(n)$ it is the case that $\langle \omega(x_1, \ldots, x_n), \sigma(y_1, \ldots, y_n)\rangle \in R$, and $\langle R, \Psi\rangle$ is locally coherent in B.

Recall that $\langle R, \Psi\rangle$ is locally coherent in A if and only if it is locally coherent in B and that $\langle R, \Psi\rangle$ is coherent if it is locally coherent in A and in B. Therefore, if Ψ is a partial function or Ψ^{-1} is a partial function, $\langle R, \Psi\rangle$ is coherent. □

How can we interpret the conditions that Ψ or Ψ^{-1} are partial functions?

Ψ being a partial function means that for any operator in the source concept network there is at most one corresponding transition in the environment. Two different transitions in the environment are never interpreted as the same operator in the concept network. Thus, there is no ambiguous operator in the source.

Conversely, Ψ^{-1} being a partial function means that for any transition in the environment there is at most one corresponding operator in the source concept network. Consequently, there is no pair of synonymous operators in Ω. For Spinner's cognitive relation, we considered in the example above, this is clearly not the case because the transition $_{\Uparrow}$ can be interpreted as r and l and $\Uparrow$ can represent both rr and ll.

Summarizing the results of this subsection, we can say that the idea that bisimulations and coherent cognitive relations express similar things, can be justified on a formal level:

1. A coherent cognitive relation is a bisimulation between the relevant subset of the source concept network and the environment.
2. Given a cognitive relation $\langle R, \Psi \rangle$ that is a bisimulation, $\langle R, \Psi \rangle$ is coherent if Ψ or Ψ^{-1} are partial functions.

5.2 Formal Analysis of the Analogical Relation that is Induced by the Algorithm HDTP-A

As explained in Section 3.2, the algorithm HDTP-A outputs a generalized theory and modifies the initial source and target theories. For a formal characterization of the analogical relation, it is important to examine these modified theories, $Th_S^{A_h}$ and $Th_T^{A_h}$, more closely because they build the domain and codomain of the analogical relation.

HDTP-A selects all elements of the initial source theory that have been anti-unified with corresponding elements of the target or that have been transferred to the target. It is this set of terms and formulas that builds the domain of the analogical relation. It represents the subset of the source that is relevant for an interpretation of the analogy in the way as it is indicated by the analogical relation. In other words, the resulting modified source theory $Th_S^{A_h}$ is exactly that subset of the initial source theory that has counterparts in the target theory.

The initial target theory is also modified; after finding corresponding pairs of source and target axioms and generalizing them, remaining axioms of the source are transferred to the target as long as this does not result in inconsistencies. It is this stage of the algorithm that models the process of inductive inference. A transfer results in an extension of the target theory, and therefore the modified target theory, $Th_T^{A_h}$, contains at least as many entities as the initial one. In the following, we will assume that all elements of the initial target domain have been generalized with an element of the source[2]. Consequently, the analogical relation R is surjective, i.e. the analogy covers the whole target domain. Especially if we consider predictive analogies, this assumption seems reasonable because an analogy that does not cover the whole source domain seems quite hard to understand. The degree of structural overlap in analogies that cover the whole target domain is referred to as *target exhaustiveness* [12].

[2]This is the case if in the algorithm HDTP-A, axioms from the target are selected first.

5.2.1 Description of the Analogical Relation with Indurkhya's Terms

In order to compare the analogical relation to a coherent cognitive relation, we will first consider the properties of the analogical relation described in the terms Indurkhya uses for characterizing cognitive models.
As explained above, the analogical relation R is defined on the whole modified source theory, $Th_S^{A_h}$; this corresponds to a cognitive model with a cognitive relation defined on the whole source concept network. In Indurkhya's framework, this property of a cognitive model is referred to as *full*.
Furthermore, if axioms from the target are selected first in the algorithm, the analogical relation covers the whole target domain. The corresponding cognitive model in Indurkhya's framework is called *complete*. Without further investigation, it is not possible to determine whether the analogical relation also has other properties like the ones described in Definition 5.

When comparing HDTP and Indurkhya's approach, a central question is whether the analogical relation and the coherent cognitive relation show similar characteristics.

5.2.2 Does the Analogical Relation Satisfy the Coherency Condition?

At this point, it is important to note that investigating this question will also reveal insights into the similarity of the analogical relation and bisimulation because of the results of Section 5.1. Proposition 1 says that a coherent cognitive relation is a bisimulation if each operator of the source is related to one of the target. Obviously, if the cognitive model is full, the coherent cognitive relation is a bisimulation. As described above, the analogical relation between the modified theories corresponds to a cognitive relation in a cognitive model that is complete and full. That indicates that if it can be shown that the analogical relation is coherent, it follows that it is a bisimulation.

The general idea of the coherency condition for a cognitive relation is that whenever one takes a pair of n-tuples that is in R and a pair of n-ary operators that is in Ψ, applying the operators to the respective n-tuples results in a pair of objects that is in R as well. For analogies, it means that applying analogous operators to analogous objects has to result in analogous objects.

Here are two of the major differences between the analogical relation and the cognitive relation that make it difficult to formulate a coherency condition in the HDTP framework:

1. In Indurkhya's framework, operators in the source concept network and transitions in the environment that are related via Ψ have to be of the same arity. In HDTP, this is clearly not the case because terms containing functions or predicates of different arities can also be anti-unified.

2. Indurkhya's coherency condition expresses an interrelationship of the operator level and the object level of the cognitive relation. In HDTP, the analogical relation is a relation between terms and formulas. There is no clearly specified restriction of this relation that corresponds to Ψ, the operator level of the cognitive relation.

Given these facts, it is obvious that a corresponding coherency condition for HDTP cannot be formulated in a straightforward way because some of the properties upon which the coherency condition relies are not clearly specified for the analogical relation in HDTP. Nevertheless, we can specify cases where the problems mentioned above do not arise.
If we restrict our investigation to cases where the function symbols and predicates in anti-unified terms are identical, the following result can be shown, which expresses an idea similar to coherency: Applying identical functions or predicates to analogous terms or formulas results in a pair of analogous items.
In the following, only a purely syntactic restriction of the analogical relation will be considered[3] because only the syntactic level is relevant if we want to compare the analogical relation to a coherent cognitive relation and to bisimulation, which also describe purely syntactic constraints.

Proposition 3. [4]

(i) Given $s_1, \ldots, s_n \in Term_S$, $t_1, \ldots, t_n \in Term_T$ *with* $\langle s_1, t_1\rangle, \ldots, \langle s_n, t_n\rangle \in R$, *it holds that:*

- *(a) For any n-ary function* f *defined on* $Term_S$ *and* $Term_T$*:* $\langle f(s_1, \ldots, s_n), f(t_1, \ldots, t_n)\rangle \in R$
- *(b)* $\langle (s_i = s_j), (t_i = t_j)\rangle \in R$
- *(c) For any n-ary predicate* P *defined on* $Term_S$ *and* $Term_T$*:* $\langle P(s_1, \ldots, s_n), P(t_1, \ldots, t_n)\rangle \in R$

[3] For a restriction of the analogical relation to the syntactic level, the third condition in Definition11 is not relevant.

[4] Note that in case two identical variables occur in two terms and the variables should not be mapped to identical terms, new variables can be introduced such that identical variables are always mapped to identical terms.

(*ii*) *Given the formulas* $\phi, \phi' \in Th_S^{A_h}, \psi, \psi' \in Th_T^{A_h}$ *with* $\langle \phi, \psi \rangle, \langle \phi', \psi' \rangle \in R$*, it holds that:*

(*a*) $\langle (\phi \wedge \phi'), (\psi \wedge \psi') \rangle \in R$

(*b*) $\langle \neg\phi, \neg\psi \rangle \in R$

(*c*) $\langle (\phi \rightarrow \phi'), (\psi \rightarrow \psi') \rangle \in R$

Proof.

(*i*) (*a*) Since $\langle s_1, t_1 \rangle, \ldots, \langle s_n, t_n \rangle \in R$, it holds that
$E_T \vdash s_i\theta_1^{-1}\theta_2 = t_i$ and $E_S \vdash t_i\theta_2^{-1}\theta_1 = s_i, \forall i \in \{1, \ldots, n\}$.
It follows that
$E_S \vdash f(s_1, \ldots, s_n) = f(t_1\theta_2^{-1}\theta_1, \ldots, t_n\theta_2^{-1}\theta_1) = f(t_1, \ldots, t_n)\theta_2^{-1}\theta_1$.
Analogously, $E_T \vdash f(t_1, \ldots, t_n) = f(s_1\theta_1^{-1}\theta_2, \ldots, s_n\theta_1^{-1}\theta_2)$
$= f(s_1, \ldots, s_n)\theta_1^{-1}\theta_2$.
Hence, $\langle f(s_1, \ldots, s_n), f(t_1, \ldots, t_n) \rangle \in R$.

(*b*) From $\langle s_1, t_1 \rangle, \ldots, \langle s_n, t_n \rangle \in R$ it follows that
$Th_S^{A_h} \cup E_S \vdash (s_i = s_j) \leftrightarrow (t_i\theta_2^{-1}\theta_1 = t_j\theta_2^{-1}\theta_1) \leftrightarrow (t_i = t_j)\theta_2^{-1}\theta_1$ and
$Th_T^{A_h} \cup E_T \vdash (t_i = t_j) \leftrightarrow (s_i\theta_1^{-1}\theta_2 = s_j\theta_1^{-1}\theta_2) \leftrightarrow (s_i = s_j)\theta_1^{-1}\theta_2$.
Thus, $\langle (s_i = s_j), (t_i = t_j) \rangle \in R$.

(*c*) $\langle s_1, t_1 \rangle, \ldots, \langle s_n, t_n \rangle \in R$ implies that
$Th_S^{A_h} \cup E_S \vdash P(s_1, \ldots, s_n) \leftrightarrow$
$P(t_1\theta_2^{-1}\theta_1, \ldots, t_n\theta_2^{-1}\theta_1) \leftrightarrow P(t_1, \ldots, t_n)\theta_2^{-1}\theta_1$ and
$Th_T^{A_h} \cup E_T \vdash P(t_1, \ldots, t_n) \leftrightarrow P(s_1\theta_1^{-1}\theta_2, \ldots, s_n\theta_1^{-1}\theta_2) \leftrightarrow$
$P(s_1, \ldots, s_n)\theta_1^{-1}\theta_2$.
Therefore, $\langle P(s_1, \ldots, s_n), P(t_1, \ldots, t_n) \rangle \in R$.

(*ii*) (*a*) Since $\langle \phi, \psi \rangle \in R$ and $\langle \phi', \psi' \rangle \in R$, it holds that
$Th_S^{A_h} \cup E_S \vdash (\phi \wedge \phi') \leftrightarrow (\psi\theta_2^{-1}\theta_1 \wedge \psi'\theta_2^{-1}\theta_1) \leftrightarrow (\psi \wedge \psi')\theta_2^{-1}\theta_1$
and $Th_T^{A_h} \cup E_T \vdash (\psi \wedge \psi') \leftrightarrow (\phi\theta_1^{-1}\theta_2 \wedge \phi'\theta_1^{-1}\theta_2) \leftrightarrow (\phi \wedge \phi')\theta_1^{-1}\theta_2$.
Therefore, $\langle (\phi \wedge \phi'), (\psi \wedge \psi') \rangle \in R$.

(*b*) From $\langle \phi, \psi \rangle \in R$ it follows that
$Th_S^{A_h} \cup E_S \vdash \neg\phi \leftrightarrow \neg(\psi\theta_2^{-1}\theta_1) \leftrightarrow (\neg\psi)\theta_2^{-1}\theta_1$ and
$Th_T^{A_h} \cup E_T \vdash \neg\psi \leftrightarrow \neg(\phi\theta_1^{-1}\theta_2) \leftrightarrow (\neg\phi)\theta_1^{-1}\theta_2$.
Thus, $\langle \neg\phi, \neg\psi \rangle \in R$.

(*c*) $\langle \phi, \psi \rangle \in R$ and $\langle \phi', \psi' \rangle \in R$ implies that
$Th_S^{A_h} \cup E_S \vdash (\phi \rightarrow \phi') \leftrightarrow (\psi\theta_2^{-1}\theta_1 \rightarrow \psi'\theta_2^{-1}\theta_1) \leftrightarrow (\psi \rightarrow \psi')\theta_2^{-1}\theta_1$
and $Th_T^{A_h} \cup E_T \vdash (\psi \rightarrow \psi') \leftrightarrow (\phi\theta_1^{-1}\theta_2 \rightarrow \phi'\theta_1^{-1}\theta_2) \leftrightarrow (\phi \rightarrow \phi')\theta_1^{-1}\theta_2$.
Hence, $\langle (\phi \rightarrow \phi'), (\psi \rightarrow \psi') \rangle \in R$. □

Proposition 3 shows that if only terms with identical functions and predicates are anti-unified, the analogical relation, induced by HDTP-A, is coherent and a bisimulation. Given a pair of n-tuples that is in R, applying the same function or predicate to them results in a pair of items that is in R as well. In other words, given an pair of n-tuples of source and target that is in R, every function, predicate and logical operator that can be applied to one of them can be applied to the other one as well such that the results are in R, too.

6 Conclusions and Future Work

6.1 Conclusions

The objective of this thesis was to compare the analogical relation in HDTP and the cognitive relation in Indurkhya's cognitive models to each other with respect to their similarity to bisimulation.
An overview of some computational approaches to analogies was given, concentrating on the formal concepts and the central relations that establish and describe structural similarities of source and target domain. Indurkhya's cognitive models and heuristic-driven theory projection, which are both based on algebraic frameworks, were investigated in more detail. Furthermore, different variations of bisimulations in several domains were presented and a connection between bisimulations and analogies was established. For this purpose, a notion of bisimulation in the framework of Indurkhya's cognitive models was introduced that is based on the idea of a generalized bisimulation with respect to a relation.

The coherent cognitive relation in Indurkhya's cognitive models was compared to this notion of bisimulation. It was shown that a coherent cognitive relation is a bisimulation between the relevant subset of the source concept network and the environment, or in the case of analogies, between the relevant subset of the source concept network and the target concept network. Moreover, it was shown that a bisimulation in Indurkhya's framework is coherent if the relation between the operators is a partial function or if its inverse is a partial function.
The analogical relation, which models the connection between source and target domain in HDTP, was compared to the cognitive relation in Indurkhya's framework. Cases were specified where the analogical relation satisfies Indurkhya's coherency condition and is also a bisimulation: If the functions and predicates in anti-unified terms are identical, the analogical relation in HDTP is coherent and a bisimulation.

Furthermore, for the general case where functions and predicates need not be identical, the results of the comparison of the coherent cognitive relation and bisimulation imply that if it can be shown that the analogical relation is coherent, it is also a bisimulation. This is due to the fact that the analogical relation in HDTP is defined on the whole modified source theory and covers the whole modified target theory.

If anti-unified functions and predicates are not identical, it cannot be seen easily whether the analogical relation satisfies the coherency condition and whether it is a bisimulation because some of the properties, the notions of coherency and bisimulation are based on, are not clearly specified for the analogical relation in HDTP.
In HDTP, there is no explicitly defined restriction of the analogical relation to function symbols and predicates that corresponds to the operator level, Ψ, of the cognitive relation. So, it is not clear how a coherency condition and a bisimulation for HDTP could be defined.
This suggests that the structural similarity of source and target that is expressed by the analogical relation in HDTP is probably weaker than bisimilarity. Of course, it has to be examined further how far it differs from bisimilarity.

As mentioned in Sections 2 and 3.1, some approaches to analogy (e.g. Indurkhya's cognitive models and Falkenhainer's Copycat) are based on the assumption that the underlying processes of analogical thinking can be seen as high level perception[1]. In [25], this is made explicit: Indurkhya uses the same formal concepts for modeling perceptual processes and analogies (cf. Section 3.1). The cognitive relation in his cognitive models is used to model both the relation between source and target in an analogy and the one between source concept network and environment, which represents the relation between an agent's conceptualization of reality, and reality itself. In the philosophy of mind, much work has been done to analyze the relations between internal representations and the things they represent. Cummins [3] describes mental representations and their relations to the entities they represent. According to him, these relations can be modeled as isomorphisms. As investigated in this thesis, a weaker form of structural equivalence as it is expressed by bisimulations might be more appropriate. At this point, it remains to analyze more precisely what kind of structural similarity is expressed by the analogical relation in HDTP. Then one would be able to determine whether this similarity is close to bisimilarity or if it is another (probably weaker) form of structural similarity that might be appropriate for describing the underlying principles of other cognitive mechanisms as well.

6.2 Future Work

Following the ideas developed in Subsection 5.2, future work would consist of a further investigation of the analogical relation in HDTP and its similarity to a coherent cognitive relation and to bisimulation. The next step would be to concentrate on the more general case where different functions and predicates of different arities are

[1]This view has also been criticized [10].

generalized. Some of the problems that would have to be solved are mentioned in Subsection 5.2.2.
For dealing with the fact that in HDTP analogous functions and predicates need not be of the same arity, a notion could be introduced that describes how an n-tuple can be related to an m-tuple.

Definition 28. *Given the analogical relation R, induced by HDTP-A and the term algebras $Term_S$ and $Term_T$ of source and target respectively, let $\vec{s} = \langle s_1, \ldots, s_n \rangle$, with $s_i \in Term_S$, and $\vec{t} = \langle t_1, \ldots, t_m \rangle$, with $t_j \in Term_T$.*
We say that $\langle \vec{s}, \vec{t} \rangle \in R$ if and only if
$\forall s_i \in \{s_1, \ldots, s_n\} \exists t_j \in \{t_1, \ldots, t_m\}$ *with* $\langle s_i, t_j \rangle \in R$ *and* $\forall t_j \in \{t_1, \ldots, t_m\} \exists s_i \in \{s_1, \ldots, s_n\}$ *with* $\langle s_i, t_j \rangle \in R$.

This could be used for a more general condition for coherency. One of the differences between the analogical relation and the cognitive relations is that, as opposed to Indurkhya's theory, in HDTP there is no formal characterization of analogous predicates or function symbols. The next step would be to define a restriction of the analogical relation for function symbols and predicates. It is not clear how this can be done because usually relations between functions are defined as relations between the results of applying the functions to arguments, e.g. for functions $f, g : \mathbb{R} \to \mathbb{R}$, $f > g := f(x) > g(x) \forall x \in \mathbb{R}$.
In the case of HDTP, analogous functions can have different domains and therefore a more complex formalization of such a restriction is needed. One possibility would be to define the restriction of the analogical relation for functions as follows[2]:

Definition 29. *Given the analogical relation R induced by HDTP-A, the restriction of R to functions $R \restriction_{Func} \subseteq Func_{\Sigma_{Th_S^{Ah}}} \times Func_{\Sigma_{Th_T^{Ah}}}$ is a set of pairs of functions $\langle f, g \rangle$ with $f : (Term_S)^n \to Term_S$, $g : (Term_T)^m \to Term_T$ such that it holds that $\langle f, g \rangle \in R \restriction_{Func}$ if and only if $\exists \vec{s}, \vec{t}$, with $\vec{s} = \langle s_1, \ldots, s_n \rangle$, $s_i \in Term_S$ and $\vec{t} = \langle t_1, \ldots, t_m \rangle$, $t_j \in Term_T$ such that $\langle \vec{s}, \vec{t} \rangle \in R$ and $\langle f(s_1, \ldots, s_n), g(t_1, \ldots, t_m) \rangle \in R$.*

Using this restriction, a generalized notion of coherency can be formulated:

Definition 30. *Given term algebras $Term_S$ and $Term_T$ of source and target respectively, the analogical relation R, induced by HDTP-A, is called* ***coherent*** *only if it holds that:*
If $\vec{s} = \langle s_1, \ldots, s_n \rangle$, with $s_i \in Term_S$ and $\vec{t} = \langle t_1, \ldots, t_m \rangle$, with $t_j \in Term_T$ and $\langle \vec{s}, \vec{t} \rangle \in R$ are given, then for any n-ary function $f : (Term_S)^n \to Term_S$ and for any m-ary function $g : (Term_T)^m \to Term_T$ with $\langle f, g \rangle \in R \restriction_{Func}$, it is the case that $\langle f(s_1, \ldots, s_n), g(t_1, \ldots, t_m) \rangle \in R$.

[2] A restriction of R to predicates could be defined analogously.

Of course, that is only a suggestion and there might be more appropriate ways to examine the analogical relation with respect to its similarity to the coherent cognitive relation and to bisimulation.

Bibliography

[1] P. Aczel. Non-well-founded sets. volume 14 of *Lecture Notes*. Center for the Study of Language and Information, Stanford, CA, 1988.

[2] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Number 53 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, UK, 2001.

[3] R. Cummins. *Representations, Targets, and Attitudes*. MIT Press, Cambridge, MA, 1996.

[4] M. Dastani, B. Indurkhya, and R. Scha. An algebraic approach to modeling analogical projection in pattern perception. In *Proceedings of Mind II: Computational Models of Creative Cognition*, pages 1–15, Dublin, 1997.

[5] M. Dastani, B. Indurkhya, and R. Scha. Modeling analogical projection based on pattern perception. In *Proceedings of Evolutionary Language Engeneering, Learning to Behave*, Belgium, 2000.

[6] B. Falkenhainer. A unified approach to explanation and theory formation. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufmann, San Mateo, CA, 1990.

[7] B. Falkenhainer, K. D. Forbus, and D. Gentner. The structure-mapping engine. In *Proceedings of AAAI-86*, pages 272–277, Philadelphia, PA, 1986.

[8] B. Falkenhainer, K. D. Forbus, and D. Gentner. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63, 1989.

[9] K. D. Forbus. Exploring analogy in the large. In Gentner et al. [16], pages 21–58.

[10] K. D. Forbus, D. Gentner, A. B. Markman, and R. W. Ferguson. Analogy just looks like high level perception: why a domain-general approach to analogical mapping is right. *JETAI*, 10(2):231–257, 1998.

[11] V. C. Galpin. *Equivalence Semantics for Concurrency: Comparison and Application*. PhD thesis, University of Edinburgh, UK, 1998.

[12] D. Gentner. The strucuture of analogical models in science. BBN Tech. report 4451, Bolt Beranek and Newman Inc., Cambridge, MA, 1980.

[13] D. Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[14] D. Gentner, B. Bowdle, P. Wolff, and C. Boronat. Metaphor is like analogy. In Gentner et al. [16], pages 199–254.

[15] D. Gentner and K. D. Forbus. MAC/FAC: A model of similarity-based retrieval. In *Proceedings of the 13th Cognitive Science Conference*, pages 504–509, Chicago, 1991. Erlbaum, Hillsdale.

[16] D. Gentner, K. J. Holyoak, and B. N. Kokinov, editors. *The Analogical Mind, Perspectives from Cognitive Science*. MIT Press, Cambridge, MA, 2001.

[17] R. Goldblatt. Mathematical modal logic: a view of its evolution. *Journal of Applied Logic*, 1(5-6):309–392, 2003.

[18] H. P. Gumm. Universelle Coalgebra. In T. Ihringer, editor, *Allgemeine Algebra*, volume 10 of *Berliner Studienreihe zur Mathematik*, pages 155–207. Heldermann Verlag, Lemgo, Germany, 2003.

[19] H. Gust, K.-U. Kühnberger, and U. Schmid. Anti-unification of axiomatic systems, http://www.cogsci.uos.de/∼helmar/analogy1.ps/, Technical Report, 2003.

[20] H. Gust, K.-U. Kühnberger, and U. Schmid. Metaphors and heuristic-driven theory projection (HDTP). Submitted to Theoretical Computer Science, 2004.

[21] M. Hennessy and R. Milner. On observing nondeterminism and concurrency. *Journal of the Association for Computing Machinery*, 32:137–162, 1985.

[22] D. R. Hofstadter and M. Mitchell. The copycat project: A model of mental fluidity and analogy-making. Technical Report CRCC-58-1992, Center for Research on Concepts and Cognition, Indiana University, Bloomington, Ind., 1992.

[23] K. J. Holyoak and J. E. Hummel. LISA: A computational model of analogical inference and schema induction. In *Proceedings of the eighteenth annual conference of the Cognitive Science Society*, pages 352 –357, Hillsdale, NJ, 1996. Erlbaum.

[24] K. J. Holyoak and J. E. Hummel. Toward an understanding of analogy within a biological symbol system. In Gentner et al. [16], pages 161–195.

[25] B. Indurkhya. *Metaphor and Cognition*. Kluwer, Dordrecht, The Netherlands, 1992.

[26] R. Milner. *A Calculus of Communicating Systems*, volume 92 of *Lecture Notes in Computer Science*. Springer-Verlag Inc., New York, NY, 1980.

[27] R. Milner. *Communication and concurrency*. Prentice Hall International Ltd., Hertfordshire, UK, 1995.

[28] D. Park. Concurrency and automata on infinite sequences. In *Proceedings of the 5th GI-Conference on Theoretical Computer Science*, pages 167–183, Heidelberg, Germany, 1981. Springer.

[29] J. J. M. M. Rutten. Universal coalgebra: A theory of systems. *Theoretical Computer Science*, 249(1):3–80, 2000.

[30] U. Schmid, J. Wirth, and K. Polkehn. A closer look at structural similarity in analogical transfer. *Cognitive Science Quarterly*, 3(1):57 – 89, 2003.

[31] J. Srba. A note on game characterizations of strong and weak bisimilarity. Course material, Aalborg University, 2005.

[32] C. Stirling. Bisimulation, model checking and other games, notes for Mathfit instructional meeting on games and computation. Technical report, University of Edinburgh, 1997.

[33] C. Stirling. The joys of bisimulation. *Lecture Notes in Computer Science*, 1450:142–151, 1998.

[34] D. Taubner. *Finite Representations of CCS and TCSP Programs by Automata and Petri Nets*, volume 369 of *Lecture Notes in Computer Science*. Springer, New York, NY, 1989.

[35] W. Thomas. On the Ehrenfeucht-Fraïssé game in theoretical computer science (extended abstract). In *TAPSOFT '93: Proceedings of the International Joint Conference CAAP/FASE on Theory and Practice of Software Development*, pages 559–568, 1993.

[36] J. van Benthem. *Modal correspondence theory*. PhD thesis, Mathematisch Instituut & Instituut voor Grondslagenonderzoek, University of Amsterdam, 1976.

[37] P. A. van der Helm and E. L. J. Leeuwenberg. Accessibility: A criterion for regularity and hierarchy in visual patterns code. *Journal of Mathematical Psychology*, 35:151–213, 1991.

[38] R. J. van Glabbeek. The linear time: Branching time spectrum (extended abstract). In J. C. M. Baeten and J. W. Klop, editors, *CONCUR'90: Theories of Concurrency - Unification and Extension (Proc.)*, pages 278–297. Springer, Heidelberg, Germany, 1990.

[39] U. Wagner. Combinatorically restricted higher order anti-unification – an application to programming by analogy. Diploma thesis, Technical University of Berlin, 2002.

I hereby confirm that I wrote this thesis independently and that I have not made use of any other resources or means than those indicated.

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Osnabrück, August 2005