Typicality:
A computational account

Tomáš Mikula

Dissertation Thesis

Department of Computer Science
Faculty of Science
Palacký University Olomouc
2024

Author
Tomáš Mikula
Department of Computer Science
Faculty of Science
Palacký University Olomouc
17. listopadu 12
CZ–771 46 Olomouc
Czech Republic

Supervisor
prof. RNDr. Radim Bělohlávek, DSc.

Keywords
typicality, concept, similarity, psychology of concept, formal concept analysis

Declaration
Hereby, I declare that the thesis is my original work. The Grammarly tool was used to make language corrections.

This thesis is based on the outcomes of joint scientific research with my supervisor Radim Bělohlávek.

Tomáš Mikula

Abstract – Typicality is one of the most influential phenomena in the psychology of concept. This work examines its computational account to find the formalization of this phenomenon. We propose several approaches to typicality, investigate their mutual relationship, and empirically test them using two datasets. We demonstrate that the Rosch and Mervis' scheme and our extended version can be reinterpreted within a similarity-based scheme through simple scaling. Subsequently, we explore enhancements to this formalization. Our novel approach, which incorporates attribute weights considering the entire category domain, significantly improves the Rosch and Mervis' scheme. These findings not only help to shine a new light on the phenomenon of typicality but also contribute to the field of data science by introducing newly proposed formulas for calculating the graded structure of categories.

Acknowledgements

I would like to express my gratitude to my supervisor prof. RNDr. Radim Bělohlávek, DSc. for sharing his fascination with cognitive psychology and his valuable guidance and inspiration, which made this PhD thesis possible. Special thanks go to all the researchers who spent their lives pursuing the fundamental questions not only of human cognition and made our research possible.

I am also deeply grateful to RNDr. Martin Trnečka, Ph.D., who motivated me to pursue the PhD degree. Professor Gert Storms for having patience with me during the long discussions during my research stay. Other PhD students of the Department of Computer Science, for sharing their experiences. Last but not least, to Gabriela and our dog, that helped me overcome hard times, especially during the COVID-19 pandemic.

This thesis is supported by grant No. IGA_PrF_2024_024 of IGA of Palacký University Olomouc.

Contents

Preface

Typicality is one of the most influential phenomena in the psychology of concept. For most people, it is natural to think that a sparrow is a better example of a bird than a penguin. Not only can we measure its influence across many behavior observations, but the paradigms and theories shifted because of typicality. Rosch and Mervis (1975) groundbreaking work opened the path for new theories and research in human cognition. Our goal is to find a formalization of this phenomenon to enrich already existing data analysis frameworks and ultimately broaden the understanding of the typicality phenomenon in general.

The thesis consists of four main chapters. Chapter 1 provides a summary of theories of concept, which suits as a ground for Chapter 2 describing the typicality phenomenon. Chapters 3 and 4 provide a summary of theoretical and experimental results of joint research work with my supervisor Radim Bělohlávek. Details are presented in five research papers, which are available as appendices A, B, C, D, and E. The first, second, and third papers are focused on formalization and experimental evaluation of typicality. The fourth paper compares many similarity measures with human similarity judgments, and the fifth paper examines the viability of formal concept analysis framework as a model of human concepts.

Note that this text is meant to accompany the papers mentioned above. Thus, readers are encouraged to study these papers in detail after reading the corresponding chapters. Readers familiar with the human conceptual system and typicality phenomenon can skip Chapter 1 and Chapter 2.

Chapter 1

The human conceptual system

The importance of concepts to everyday life is indisputable. Without concepts, one would starve surrounded by tomatoes because they had never experienced those particular tomatoes before (G. Murphy, 2004, p. 3). We access knowledge about tomatoes every time we categorize a novel object as a tomato, infer that green tomatoes are probably not tasteful, understand the meaning of the word "tomato" in a sentence, or make an analogy between tomatoes and other objects (Barsalou, 2012; Machery, 2009).

The authors wrote about concepts in diverse ways. George Murphy used the metaphor of mental glue (G. Murphy, 2004, p. 1). Barsalou described a concept as the basic unit of knowledge but also proposed a theory of the human conceptual system without an explicit definition of concept (Barsalou, 2012; Barsalou, Kyle Simmons, et al., 2003). Laurence and Margolis discussed concepts as a fundamental construct in theories of the mind (Margolis and Laurence, 1999, p. 3). Machery formulated a definition of the theoretical term *concept* within cognitive psychology as follows:

> A concept of x is a body of knowledge about x that is stored in long term memory and that is used by default in the processes underlying most, if not all, higher cognitive competences when these processes result in judgments about x. (Machery, 2009, p. 12)

To fully understand this definition, we have to introduce two concepts: long-term memory and higher cognitive competencies. First, long-term memory stores a wide variety of information (e.g., what tomatoes are, the last time we ate a tomato, how to cut a tomato in half). This information is usually stored for several decades, compared to short-term memory, which can store information for a few seconds[1] (Eysenck and Keane, 2020, p. 240). Second, with a little bit of controversy, we can divide cognitive competencies (e.g., vision, motor planning, categorization, induction) into higher and lower ones:

> For present purposes, suffice it to say that the lower cognitive competencies encompass our perceptual competencies and our motor competencies, although

[1] The division to long-term and short-term memory comes from the multi-store model, which focuses primarily on their differences (Atkinson and Shiffrin, 1968). It is worth mentioning that this approach is not the only one. The unitary-store model focuses more on their similarities since they are mutually affected (Eysenck and Keane, 2020, p. 244).

> the last stages of perception, particularly the categorization of what is perceived, belong to higher cognition. (Machery, 2009, p. 8)

The fact that concepts are used in a wide variety of cognitive tasks was described as *promiscuity of concepts* (Machery, 2009, p. 16). This promiscuity is likely even more significant, as we will see in the hypothesis of the motor system being tied to the conceptual system (Section 1.2.6).

It is worth mentioning that this definition does not say how and in which form this knowledge is stored nor how and when it is accessed. Machery wrote explicitly that the given definition is not a theory of concepts (we will discuss these later in this chapter), and he tried to stay as general as possible.

Research around concepts is getting more complex every year. An increasing amount of research combines multiple approaches (e.g., cognitive neuroscience and computational cognitive science) to examine human cognition (Eysenck and Keane, 2020, p. 33). Thus, we will not describe every aspect of the concept phenomenon. Instead, we will demonstrate the complexity of the human conceptual system, which can be easily overlooked in everyday life. George Murphy concluded the introductory section of the Big Book of Concepts beautifully:

> My guess is that in twenty more years, it will be impossible to write a single-authored book that covers the same ground, because there will be more research than any one author and volume can handle. (G. Murphy, 2004, p. 8)

1.1 Possible pitfalls

Before we move to specific theories of concepts, we must address multiple pitfalls of research on the human conceptual system. This may sound predominantly negative, but it is essential to have these in mind early on.

Blunden wrote about two assumptions that are central to many theories of concepts. The first assumption was called *cartesian dualism* (Blunden, 2012, p. 14). According to Blunden, many psychologists understand concepts as mental representations of material objects. This assumption led us to the state where we have to distinguish concepts (mental representation or body of knowledge) from *categories* (groups of material objects or class of objects) (G. Murphy, 2004, p. 5). The reasonable assumption quickly becomes problematic since psychologists often use these terms interchangeably (Machery, 2009). Murphy emphasized that being too fussy about separating these two notions leads to problems with a slight advantage in clarity (G. Murphy, 2004, p. 5). The category notion is also often used for classes of three-dimensional, medium-sized objects such as animals and artifacts (Machery, 2009, p. 12).

The main tenet of the second assumption was that the objective world is arbitrary, atomistic, and composed of individual entities (Blunden, 2012, p. 14). Therefore, every entity can be fully described by its attributes (or features). On the contrary, this assumption is not always observable in laboratory experiments where subjects expect some other essence:

> A dog is a member of the species dog because it was born of a dog, not because it is like other dogs. (Margolis and Laurence, 1999, p. 530)

Humans acquire concepts through the social experience of everyday life, and isolating subjects during laboratory experiments does not have to yield relevant results (Blunden, 2012, p. 14). Nevertheless, formulating hypotheses, bringing the appropriate number of subjects into the laboratory, and testing formulated theories was the primary approach during the cognitive revolution (Griffiths, 2015). Griffiths (2015) argued that the world has changed, and psychologists are no longer people with the most data about human behavior. Computer scientists can access extensive databases that reflect human cognition more than behavioral data from laboratory experiments. Griffiths urged the development of new experimental paradigms that are suitable for a large number of respondents. Murphy proposed that essential and surprising discoveries are still being made by studying real-world concepts in greater detail (G. Murphy, 2004, p. 7).

One could say that neuroimaging technology like functional magnetic resonance imaging (fMRI)[2] would solve these issues. The ability to monitor brain activity often leads to even more significant isolation since subjects must be seated inside a large tunnel-shaped device that emits loud sounds. Despite that, it must be stated that the contribution of fMRI and other neuroimaging technology to human cognition research is enormous.

Even when we overcome these issues, we must be aware that the study of categorization only tells us a little about the concepts in general. That was pointed out by Solomon, Medin, and E. Lynch (1999):

> Until recently, the study of concepts has largely been the study of categorization. However, categorization is only one conceptual function among several. We argue that concepts cannot be understood sufficiently through the study of categorization, or any other function, in isolation for two important reasons. (Solomon, Medin, and E. Lynch, 1999)

The focus on categorization was useful mainly at the beginning of experimental research of concepts. However, this focus possibly led to paying less attention to other essential aspects of concepts (Machery, 2009, p. 29). Many authors recognized these problems (G. Murphy, 2004). As mentioned earlier, cognitive psychologists are often part of multidisciplinary research teams, allowing us to understand these complexities more broadly.

Last but not least, an issue inside experimental cognitive psychology (and science in general) is the reproducibility of past experiments[3] (Goodman, Fanelli, and Ioannidis, 2016). Only 50% (21 of 42) of findings in cognitive psychology were successfully reproduced (Eysenck and Keane, 2020; Open Science Collaboration, 2015). Stanley and Spence

[2]The fMRI uses a scanner with a huge magnet whose weight can be up to 11 tons. Oxyhaemoglobin is converted into deoxyhemoglobin when neurons consume oxygen, leading to distortions in the local magnetic field. These distortions can be assessed by fMRI, which measures the concentration of deoxyhemoglobin in the blood. These changes represent the neural activity of the brain. (Eysenck and Keane, 2020)

[3]It is worth mentioning that deep learning (part of computer science) suffers from similar reproducibility issues.

(2014) argued that expectations of reproducibility are unreasonable. Human cognition is too complex to yield easily reproducible results. Still, it does not mean we cannot trust any experimental results. Numerous important results were replicated dozens of times, and it is clear that they remain important for future research.

1.2 Theories of concepts

The following sections will introduce multiple so-called *theories of concepts*. These theories usually describe the kind and format of the knowledge stored in concepts, how concepts are involved in cognitive processes, how they are acquired, and where they are localized in our brains (Machery, 2009, p. 20).

Modern cognitive psychology, cognitive science, and cognitive neuroscience can separate these theories into three groups (Barsalou, 2012): classic *good old fashioned artificial intelligence* (GOFAI) approaches, which benefit from manipulation with symbols; *statistical approach*, which implements dynamic and situated conceptual representations; *simulation/situated approaches* that ground conceptual knowledge in modality-specific systems. It is worth noting that the development of artificial intelligence more or less follows similar trends, currently firmly rooted in a statistical approach looking for some progressive neuro-symbolic hybrid systems (Garcez and Lamb, 2020).

Theories can be further divided by an amodal and modal representation of concepts (Barsalou, 2012). The *amodal* approach involves the redescription of modal-specific information (e.g., visual representation of attribute “red”) into amodal representation (e.g., symbolic representation of “red” in the form of language). The modal representation is not used directly. On the other hand, *modal* approaches assume that the human conceptual system is grounded in the brain’s modality-specific systems, body, and environment. Therefore, these modal representations are directly involved and necessary.

Another possible division of theories is in terms of *stability*. Theories that assume unstable concept representations propose that the conceptual system is dynamic and situated, leading to representations tailored to the current needs of situated action (Barsalou, 2012). The stable theories assume that each concept has one *core* universal across different situations and needs.

Reading these theories in the context of possible pitfalls mentioned in the previous section is essential. In 2004, Murphy described being uneasy about the disputes in the field since much of the literature compared two particular theories, which seemed to be wrong to a greater or lesser degree (G. Murphy, 2004, p. 4). Nevertheless, these theories have discovered a fundamentally important discovery about the human conceptual system (Barsalou, 2012). Writers agreed that future theories must be some integration of the previous ones (G. Murphy, 2004, p. 488, Barsalou, 2012).

Some theories are more important to this thesis than others. Despite that, we will briefly describe the most influential ones.

1.2.1 Classical theory

According to the so-called *classical theory* (Smith and Medin, 2013), concepts are characterized by definitions similar to the dictionary. Every definition has two aspects: *necessity* and *sufficiency* (Machery, 2009; G. Murphy, 2004). The entity is a category member if it possesses all attributes from the definition (necessity). At the same time, definitional attributes are jointly sufficient to the entity being a member of a given category (sufficiency). The definition of a concept is assumed to be amodal and stable. This definition approach classifies classical theory as GOFAI theory. Hull described the core tenet of classical theory as follows:

> All of the individual experiences which require a given reaction, must contain certain characteristics which are at the same time common to all members of the group requiring this reaction and which are NOT found in any members of the groups requiring different reactions. (Hull, 1920; G. Murphy, 2004)

One of the main implications of classical theory is that every entity is either a member of the given category or not. This allows us to see concepts in the context of traditional logic (Inhelder and Jean, 1964; G. Murphy, 2004).

The main problem with this approach to concepts is that definitions inside dictionaries need to provide a realistic idea of the corresponding concepts [p. 17](Blunden, 2012). The assumption that concepts can be clearly defined was questioned by philosopher Wittgenstein (1968) relatively early in the example of concept "game". It is impossible to write a definition of games that includes all games and excludes non-game sports like hunting (G. Murphy, 2004, p. 17).

Wittgenstein's objections were evaluated experimentally (G. Murphy, 2004). Hampton (1979) asked subjects to produce features as definitions of eight categories and rate items on the scale of category membership. He concluded that experiments do not suggest the existence of attributes that provide necessary and sufficient definitions. McCloskey and Glucksberg (1978) asked subjects similar questions in two experimental sessions separated by approximately one month. They have found that within-subjects inconsistency varied across different members of categories.[4]

One would assume that the rules of games can be used as a perfect example of concepts with crystal clear definitions. The opposite is true; in 1999, Major League Baseball tried to standardize the strike zone; the interesting fact is that baseball was played successfully before that (G. Murphy, 2004, p. 19).

An even more disturbing example is the field of law. Legal practice has shown that law is fuzzier than we hoped. The fuzziness is inevitable because lawmakers cannot cover every situation the law has to address (G. Murphy, 2004, p. 19). For example, if we formulate a rule "No vehicles in the park", do we include wheelchairs? A similar situation arose when new means of transportation became popular. The rise of injuries caused by electric scooters being in the gray zone forced new law changes (Blomberg et al., 2019). The inability of classical theory to explain border cases and the graded nature of category membership ultimately led to its fall.

The logical aspect of classical theory provides a natural way of ordering concepts hierarchically via transitivity (G. Murphy, 2004, p. 27): If all As are Bs, and Bs are Cs, then all As must be Cs. Hampton (1982) described the *failure of transistivity.* Subjects

[4]We are omitting details here; inconsistency varied across typicality range of category members; typicality account will be explained in Chapter 2.

judged whether an armchair is a chair and, at the same time, a piece of furniture. Then they judge if a chair is also furniture. Transitivity easily fails in examples like the car seat being a chair, a chair being a piece of furniture, but a car seat is not a piece of furniture.

These problems led to the conclusion that a dictionary-like structure does not represent concepts. Blunden argued that this is not a problem of psychology, but it is the nature of the concepts themselves:

> Concepts are not pigeonholes and concepts which conformed to expectations of these researchers would be very poor concepts ... Concepts have a content which is objective, and insofar as concepts reflect the material world, they will be inconsistent, unstable and contradictory. (Blunden, 2012, p. 18)

The efforts to eliminate these contradictions instead of accepting them as the core idea of concepts could lead us to miss the point completely:

> This was the result when recent investigators in the Psychology of Concepts wrongly interpreted the failure of *their interpretation* of the Classical Theory as a problem of *psychology* ... If we let go of the idea that Cognitive Psychology is some variety of Psychology, and instead regard Cognitive Psychology as a branch of Engineering Science, then all this makes abundant sense. But then surely Cognitive Psychology loses its very *raison d'être* if it stops paying attention to what is distinctively human and unlike a machine in human behaviour? (Blunden, 2012, p. 19, 20)

On the other hand, Murphy warned about possible problems of proposing that concepts in cognitive psychology are not the "real concepts":

> Writers with a background in linguistic semantics or philosophy have come to different conclusions about what concepts really are, often with a "classical" flavor–conclusions which are not compatible with the typical results of psychology experiments.

At the same time, he warned about trusting too much to laboratory experiments:

> However, we should not make the mistake of believing too much in the concepts we make up for experiments, which probably drastically underestimate the complexity and richness of real-world concepts. (G. Murphy, 2004, p. 48)

Does this mean that we have dictionary-like structures in our heads? Probably not, since experimental evidence is strong. Nevertheless, we should not reject some ideas *only* based on laboratory experiments with object categorization.

1.2.2 Prototype theory

After a wave of criticism in the 1970s, the *prototype theory* quickly replaced the classical theory. Posner and Keele (1968, 1970), Rips, Shoben, and Smith (1973), Hampton (1979) and Smith and Medin (2013) were most influential authors of prototype theory, although its roots lie in a groundbreaking publication from Rosch and Mervis (1975) about typicality effect which will be detaily discussed in Chapter 2.

Many readers interpreted the idea of prototype theory in a way that every concept is represented by *single prototype* (e.g., the ideal exemplar). All other members are included or excluded in the concept according to how much they *resemble* the *prototype* (Machery, 2009; G. Murphy, 2004, p. 22). Instead, a prototype should be understood as (statistical) *summary representation*:

> The entire category is represented by a unified representation rather than separate representations for each member or for different classes of members. (G. Murphy, 2004, p. 42)

Representing concepts by attributes usually present in the category of exemplars is one way of understanding summary representation. The main difference to classical theory is that we do not require all exemplars to have these attributes. Each attribute can be crucial to a different degree (G. Murphy, 2004, p. 43).

Instead of matching specific formal definitions during the categorization task, the subject calculates the exemplar's similarity to the set of defining features. Exemplar then gets a weighted score for each matching attribute, which is summed and compared to *categorization criterion*, which can be understood as a categorization threshold that has to be met to exemplar being a member of the given category (G. Murphy, 2004, p. 44). This approach allows us to explain the border cases and the gradual structure of membership functions.

In comparison to classical theory where concepts are ordered in hierarchical taxonomy, the prototype theory uses the topology of distinct clusters of concepts (Blunden, 2012, p. 22). This allows us to address the transitivity problem mentioned in the previous section. A car seat is a chair, and a chair is a piece of furniture, but a car seat does not have to be a member of furniture since it is not similar in essential attributes to the concept of furniture (Tversky, 1977; G. Murphy, 2004, p. 45).

The prototype theory is still considered to be part of the GOFAI theories since it builds upon the principle of semantic memory, which contains symbols representing concepts (Barsalou, 2012) and relies heavily on an amodal and stable approach to concept representation. From this point of view, prototype theory is similar to classical theory.

1.2.3 Exemplar theory

A few years after prototype theory, the *exemplar theory* paradigm was introduced by Medin and Schaffer (1978) and Brooks (1978). The main difference to prototype theory is that people do not have one representation that describes the whole concept; instead, one's concept is represented by *exemplars* which were experienced through life (G. Murphy, 2004):

> The general idea of the context model is that classification judgments are based on the retrieval of stored exemplar information. (Medin and Schaffer, 1978)

Most exemplar approaches remain amodal by storing exemplars in memory outside the modality-specific systems. However, some exemplar models that store exemplars inside modal-specific systems exist (e.g., an exemplar of a visual category is stored as a visual memory in the visual system) (Barsalou, 2012).

Similar to prototype theory, the categorization is based on similarity. During the categorization process, similarity to all exemplars is calculated for each category (e.g., resemblance to all remembered dogs, cats, etc.). The calculated similarities are summed, and the category with the highest sum is selected. This leads us to the conclusion that exemplar theory assumes the stability of concept representation (Barsalou, 2012).

The problem of borderline cases is not present since can be explained by situations where two categories have almost identical similarities after the summation. Transitivity can be explained similarly as in the case of the prototype theory (G. Murphy, 2004). A car seat is similar to a chair exemplar in some aspects; a chair is similar to furniture in different aspects. However, car seats differ from furniture exemplars; therefore, they are not considered furniture.

Since a concept is represented by exemplars and no abstract prototypes are built, the biggest problem of exemplar theory are symbolic operations (Barsalou, 2012). On the other hand, this theory excels in categorization since detailed category information is encoded in exemplars (Barsalou, 2012).

1.2.4 Theory theory

As we stated before, most of the research was conducted inside laboratory conditions in isolation of our everyday world knowledge (G. Murphy, 2004, p. 141). These experiments were motivated by maximizing control over unknown variables. However, it is questionable if this approach led us to a better understanding of the human conceptual system (G. Murphy, 2004, p. 141).

Influence of prior knowledge known as *knowledge effect* was examined during the mid-1980s (Carey, 1987; G. L. Murphy and Medin, 1985). Knowledge effect refers to the influence of prior knowledge of real objects and events in category-learning situations. Prior knowledge can be involved in multiple parts of the conceptual system (G. Murphy, 2004, p. 146). Firstly, it can influence which features will be selected as definitory for a given object (e.g., the influence of the context of the situation). Secondly, prior knowledge can improve the learning of new features of new categories (e.g., it is harder to learn an arbitrary list of features instead of features that pose coherent structures). Lastly, it can influence categorization decisions (e.g., inference from the context).

Neither prototype theory nor exemplar theory have addressed the knowledge effect. It is mainly caused by their approach to concept representation, which is built only by experience with exemplars (G. Murphy, 2004, p. 183). That is where theory theory emerges.

The central tenet of *theory theory* is that concepts are part of our general knowledge, which are not learned in isolation; instead, they are learned in consistency as part of our complete understanding of the world (G. Murphy, 2004, p. 60). As a consequence, *naive theories* which hold casual relations between exemplars in domains are formed (Rogers and McClelland, 2004, p. 28).

The importance of the knowledge approach can be demonstrated on *goal-derived categories* (e.g., "things to eat on a diet") (G. Murphy, 2004, p. 62) or *ad-hoc categories* (e.g., "what to do when chased by a mafia") (Barsalou, 1983). The similarity of exemplar features cannot easily explain the structure of these categories. Barsalou (1985) proposed that so-called *ideals* (e.g., zero/low calories) can be a better approach.[5]

The theory theory is trying to explain only some of the concept learning since some categories cannot be based on previous knowledge (G. Murphy, 2004, p. 63) (e.g., the relevant knowledge can be accessed only after examining the exemplars of a given category). Instead, it is usually viewed as a complementary approach that can help categorization-based approaches overcome some of their limitations (e.g., taxonomic organization) (Rogers and McClelland, 2004, p. 29). One problem of theory theory is the lack of models of the knowledge approach (Machery, 2009, p. 105).

1.2.5 Connectionist theory

Categorization-based approaches to human semantic memory involve discretization of a graded set of similarity relations which lead to unwanted discarding of important information (Rumelhart, Smolensky, et al., 1986, Rogers and McClelland, 2004, p. 44). This motivated the development of models that learn graded, similarity-based generalization properties of categories (Rogers and McClelland, 2004, p. 46). Therefore, the *connectionist theory* emphasizes the distributed representation of concepts without needing GOFAI symbols inside semantic memory.

Most influential connectionist approach to semantic cognition is called *Parallel distributed processing* (PDP) (Rumelhart and McClelland, 1986) based initially on Hinton's work of general framework for storing propositional knowledge in distributed connectionist network (Hinton, 1986; Hinton and Anderson, 1989; Rogers and McClelland, 2004).

The simplified Rumelhart feedforward network model acquires distributed representations of exemplars by learning their properties in different relational contexts. The model is represented by a multi-layer network built with nonlinear processing units connected in a feed-forward manner (Rogers and McClelland, 2004, p. 58). The *Item layer* represents perceptual system (Barsalou, 2012) for an input representation of exemplars in the world (Rogers and McClelland, 2004, p. 58), the *Relation layer* can be understood as a simplified context specification, and lastly, the *Attribute layer* represents the predicted consequences following the occurrence of the exemplar in the given context.

The network models compute as follows:

> Patterns are presented by activating one unit in each of the *Item* and *Relation* layers (i.e., these activations are set to 1 and activations of all other input units are set to 0). Activation then feeds forward through the network, modulated by the connection weights. Activations are updated sequentially, layer by layer, so that first the representation layer is updated, then the hidden layer,

[5]Ideals will be discussed in the Section 2.2.2.

> then the attribute layer. To update the activation of a unit, first its net input is calculated: the sum, over each of the unit's incoming connections, of the activity of the unit sending its activation through the connection, multiplied by the value of the connection weight. The net input is then transformed into an activation according to the logistic activation function .... (Rogers and McClelland, 2004)

The initial values of connection weights are small random values. This approach ensures initial random relation input to output (Rogers and McClelland, 2004, p. 58). The introduction of groundbreaking back-propagation learning algorithm (Rumelhart, Hinton, and Williams, 1986) brings the possibility that weights of connections can be learned. That is done by presenting target values to the network model and back-propagating the error to previous layers. So, the learning process is gradual during a presentation of many training examples.

The activations triggered by a single exemplar in the input generate a distributed pattern across units in the following layers. The distributed approach to concept representation provides a natural way of generalization (Rogers and McClelland, 2004, p. 52). The weights between *Item* and *Representation*, which emerge during the learning process, represent internal concept representation (Rogers and McClelland, 2004). This concept representation can be viewed as amodal, situated, and dynamic (Barsalou, 2012).

The weakest aspect of connectionist models is the lack of symbolic operations (propositions and productivity). The *proposition* establishes a relation between individual exemplar and concept. The *productivity* underlines people's creative abilities to combine words and concepts into complex linguistic and conceptual structures (Barsalou, 2012). It is worth mentioning that some impressive progress was made in the deep learning field (Lake, Salakhutdinov, and Tenenbaum, 2015; Lake, Ullman, et al., 2017/ed; Ramesh et al., 2022).

1.2.6 Situated theory

Approach to human cognition sometimes called *simulation* or *embodiment* theory (Barsalou, 2012) proposed mainly by Barsalou, define concept as follows:

> ... a concept is a dynamical distributed system in the brain that represents a category in the environment or experience and that controls interactions with the category's instances (e.g., the concept of bicycle represents and controls interactions with bicycles). (Barsalou, 2015)

Barsalou (2015) stated that the conceptual system is involved virtually in all cognitive processes. Concepts contribute to perception during online interactions with the environment, enable categorization, support action (e.g., predicting which actions will be effective), and are an essential part of offline processing when people represent nonpresent situations (e.g., imagination, memory).

Concept theories mentioned before usually assume that concepts are abstract, detached from input (sensory) and output (motor) processes, stable across different contexts, and similar across multiple individuals (Eysenck and Keane, 2020, p. 316). This assumption led to studies where perceptual and motor systems were not accounted for (Barsalou, 2015). Barsalou (2015) argued that the so-called *sandwich model* (conceptual system sandwiched between sensory and motor processes) would never fully explain the human cognition system.

On the contrary, *grounded cognition* accounts for several aspects not included in previous theories (Barsalou, 2015). Firstly, according to this theory, cognition relies heavily on modalities (perception, action, and interception). Secondly, cognition often relies on the state of the body and psychical action. Lastly, cognition depends on the physical and social environment. From a grounded cognition point of view, cognition emerges from interaction with sensory-motor systems, the body, and the psychical and social environment. The conceptual representation is distributed across all of them. This concludes as the modal and nonmodular conceptual system (Barsalou, 2012).

According to situated theory, the *simulation* plays a vital role in the conceptual system, which establishes a brain state similar to the one that occurs while interacting with the category's exemplars (Barsalou, 2015). To a large extent, simulations remain unconscious while influencing cognition processes. When simulation becomes conscious, we experience mental imagery.

The simulations are constructed in *situated* (dynamic) manner (Barsalou, 2015). The category can be simulated in various contexts, and the most relevant simulation is selected. The category is then represented by a large set of simulations similar to exemplar theory. This approach can explain differences in concepts across multiple individuals since situations are expressed as a set of unique memories.

One might say that situated theory does not account for abstract concepts (e.g., truth, love, freedom). Barsalou, Dutriaux, and Scheepers (2018) stated that abstract concepts are experienced in concrete contexts. Therefore, they can be accounted for by situated cognition (Eysenck and Keane, 2020, p. 317).

Situated theory demonstrate the importance of modalities, context, individual goals, and situated aspects of the conceptual system alongside concepts grounded in the same neural system as perception and sensory-motor system. As we will see in the following section, the presence of a stable core has yet to be disproved (Borghesani and Piazza, 2017; Eysenck and Keane, 2020, p. 318).[6]

1.2.7 Controlled semantic cognition

Ralph et al. (2017) published an influential paper based on a decade of neurocognitive and neurocomputational research, which concluded in a new controlled semantic cognition (CSC) framework that combines some of the core ideas from connectionist theory and situated theory. The SCS framework, based on decades of research in semantic dementia and semantic aphasia, consists of two systems: *semantic representation* and *semantic control*.

[6]For detailed discussion see Borghesani and Piazza, 2017.

Semantic representation

The semantic representation system is described by *hub-and-spoke* model (Patterson, Nestor, and Rogers, 2007). Concept representation is constructed by encoding high-order relationships from different modalities and comprises two core ideas: the hub and the spokes.

The *hub* corresponds to classic approaches to conceptual systems, which propose that concepts are represented as stable context-insensitive amodal conceptual cores. Hub consists of interconnected units similar to other models in connectionist theory. Empirical and computational observations of semantic dementia impairment support the hub's existence. Patients with semantic dementia have difficulties with tasks that involve recruiting knowledge of categories across all modalities and almost any type of categories (Mayberry, Sage, and Ralph, 2011; Ralph et al., 2017). Experimental research found support for equal disruption across types of knowledge as well as category-specific impairments. To account for both findings, the *graded hub* was proposed (Ralph et al., 2017). Instead of the original hub model in which all units have an equal contribution to the semantic representation, the graded hub introduced units that contribute in a graded manner.

The *spokes* represents the brain's modal-specific regions (e.g., sound, speech). Each spoke consists of interconnected units connected to the hub. These units provide a modal part of the amodal knowledge stored in the hub. This can explain the findings of modal representation mentioned by Barsalou in the situated theory. The existence of spokes is supported by various neurocognitive research (Ralph et al., 2017). One example can be Pobric, Jefferies, and Lambon Ralph (2010) which used transcranial magnetic stimulation (TMS)[7] to interfere with the brain region which is associated with paraxis spoke (responsible for processing actions which can we make toward objects) (Eysenck and Keane, 2020) – resulting in degrading category-specific performance for nonliving items.

Together, hub and spoke provide distributed conceptual representation based on stable context invariant core and modal-specific representation in corresponding spokes.

Semantic control

The primary role of the semantic control system is manipulating semantic representation (e.g., suppressing dominant attributes), which allows different behaviors in different contexts. The existence of this system is supported by research on semantic aphasia disorder in which patients struggle with task-dependent manipulation of conceptual knowledge (Ralph et al., 2017). As we described in Section 1.2.5 and Section 1.2.6, many phenomena require concept representation to adjust dynamically. The CSC framework implements semantic control via a distributed neural network largely separated from semantic representation (Ralph et al., 2017).

The PDP model can model the hub and spokes in which the hub is represented by representation layers, the spokes by output layers, and the semantic control is modeled by the hidden layers influenced by the context nodes, which forms the relation layer (Folstein and Dieciuc, 2019).

[7]The TMS sends very brief pulses of current into the participant's brain, which inhibits processing inside the applied region (Eysenck and Keane, 2020, p. 16).

The computational model based on a recurrent neural network (RNN) was used to reverse-engineer the possible brain architecture of CSC (Jackson, Rogers, and Lambon Ralph, 2021). The authors proposed three core functions of the human semantic system by which computational models were evaluated:

> (1) It must acquire representations that capture the overall conceptual similarity structure and not merely the perceptual, motor and linguistic structures apparent within various modalities.
>
> (2) It must acquire context-independent conceptual representations from learning episodes that provide only partial, context-specific information about an item's properties.
>
> (3) It must adapt to context so as to generate only context-appropriate behaviours.
>
> (Jackson, Rogers, and Lambon Ralph, 2021)

The computational model evaluation showed that the presence of a multimodal hub allows the model to acquire richer internal representations, the deep of the RNN does not automatically yield better performance, and the shortcut connections that can skip the model's layers provide faster learning and better conceptual abstraction.

The CSC framework is a promising approach to describing the human conceptual system since it addresses multiple problems that previous theories cannot address. Nevertheless, there are still questions that must be answered. One example is the relative contributions of the hub versus spokes or representation of abstract, emotional, and social concepts (Ralph et al., 2017). Another challenge is ad-hoc categories and situational aspects of human cognition, which were addressed in situated theory in Section 1.2.6 (Folstein and Dieciuc, 2019).

1.3 Conclusions

In this chapter, we have discussed the importance of the human conceptual system and possible pitfalls during its research, as well as summarized current modeling efforts.

Research on human conceptual abilities is expected to become even more active since many other fields are interested in this area (e.g., artificial intelligence). Interdisciplinary research is becoming more valuable since cooperation across multiple fields can accelerate the research. For example, cognitive neuroscience can provide new ideas that can be tested in the field of artificial intelligence. Alternatively, existing models inside artificial intelligence should be evaluated with the criticism of cognitive neuroscience.

The wide range of theories presented in this chapter should provide the idea that the "correct answer" is still far away in the future. However, history provided an important lesson: there is usually no single best theory, and combining multiple approaches and iterative research usually brings the most interesting results.

Chapter 2

The typicality phenomenon

Why is a sparrow a better example of a bird than a penguin, or why is hockey a better example of a sport than a checker (Rosch, 1975)? The *typical* members are good examples of a category. The *atypical* exemplars are known to be members of a category but are unusual in some sense (G. Murphy, 2004, p. 22). These observations show the graded structure of category, which led to the rise of new theories of concepts previously discussed in Chapter 1. This chapter will describe the main behavioral observations and possible explanations of typicality.

2.1 Behavioral observations

The typicality effects are among the strongest and most consistent in categorization literature (G. Murphy, 2004, p. 22). We will briefly introduce some of the behavioral results.

Category judgement inconsistence

Let us recall McCloskey and Glucksberg (1978) research mentioned in Section 1.2.1. They did two experiments separated by approximately one month in which participants judged exemplar/category name pairs (e.g., apple/fruit) on a 3-point scale ("yes" as being in a category; "no" as not being in a category; "unfamiliar" for cases where participants were not familiar with the exemplar). They showed that participants changed their minds mainly in the case of exemplars who were not typical or apparent non-members of the category.

Similarly, Rips, Shoben, and Smith (1973) did an experiment where participants pressed a "yes" or "no" button in situations where they thought that exemplar is a member of a category (e.g., "a sparrow is a bird"). After, they examined the relationship between typicality and reaction times. They found that people are much slower in the case of atypical exemplars (e.g., if a chicken is a bird).

Category learning

The typicality effect can be largely observed in category learning. Rosch, Simpson, and Miller (1976) did a series of experiments on three artificial categories: dot patterns, stick figures, and letter strings. They showed that participants list the typical exemplars more often:

> For all three types of category, the rate of learning, reaction time in a category verification task, rating of typicality of instances, and order and probability of output of category members were shown to be a function of the structure of the category. (Rosch, Simpson, and Miller, 1976)

Mervis and Pani (1980) did similar experiments on different artificial stimuli, emphasizing similar category structures as natural object categories. They created 24 three-dimensional objects designed to form six categories in a graded manner. They found support for two hypotheses: categories are learned easier through exposure to good (typical) exemplars, and good exemplars are learned before the bad (atypical) ones.

Category inference

Rips (1975) found that typicality influences subjects' judgments during category inference. The subject judged the probability of new unknown diseases spreading among other animal species. When a typical species exemplar is presented (e.g., a robin), subjects assume that diseases can spread to any other bird. On the other hand, when an atypical exemplar is presented (e.g., a duck), subjects assume that diseases cannot spread to all birds. Similar results were obtained by Osherson et al. (1990).

Exemplar generation frequency

The frequency of category member generation is also related to the typicality. Mervis, Catlin, and Rosch (1976) showed a high correlation between goodness-of-example (typicality) and item dominance. Item dominance describes the frequency of generation of some exemplar for the given category as stimulus (Mervis, Catlin, and Rosch, 1976).

Linguistics

Kelly, Bock, and Keil (1986) found the influence of typicality in the structure of the sentences. The typicality influences the order of exemplars in the sentence; the most typical is usually mentioned first. The sentences where the typical exemplars are mentioned before the atypical ones are considered more natural.

Garrod and Sanford (1977) examined the time needed to read the sentence. They found that subjects spent more time reading the sentences where an atypical member of a category was presented.

2.2 Explaining typicality

We will now focus our attention on examining the possible determinants of typicality. Three approaches will be presented in more detail. Firstly, Rosch and Mervis' groundbreaking paper in which they proposed the family resemblance hypothesis (Rosch and Mervis, 1975). Secondly, Barsalou's influential work about ideals (Barsalou, 1985). Lastly, Dieciuc and Folstein's overview of the possibility that typicality can be explained by the conjunction of these two approaches (Dieciuc and Folstein, 2019; Folstein and Dieciuc, 2019).

2.2.1 Family resemblance

Rosch and Mervis (1975) did multiple experiments that examined the internal structure of categories. They proposed that the graded structure of categories could be explained by *family resemblance*:

> The basic hypothesis was that members of a category come to be viewed as prototypical of the category as a whole in (1) proportion to the extent to which they bear a family resemblance to (have attributes which overlap those of) other members of the category. Conversely, items viewed as most prototypical of one category will be those with (2) least family resemblance to or membership in other categories. (Rosch and Mervis, 1975) (added parts numbering)

The family resemblance hypothesis was examined in three types of categories. The superordinate categories, basic-level categories, and artificial categories. Before diving into the results, let us describe these types of categories in detail. The superordinate and basic level categories are two important types of categories. The categories that are higher in the conceptual hierarchy are *superordinate* to the lower-level categories (e.g., "furniture", "vehicle"); the lower-level categories are *subordinate* to the higher-level ones (e.g., "wooden upholstered chairs", "red sports cars") (G. Murphy, 2004, p. 200). The basic-level categories (e.g., "chair", "sports car") lie in between – they are not too general nor too specific. The basic-level categories are the preferred level to conceptually divide the world (G. Murphy, 2004, p. 210). This hierarchy also has an important implication for the number of shared attributes. The members of superordinate categories usually have only a few common attributes, the members of a basic-level category share significantly more attributes, and the members of subordinate categories share slightly more attributes than members of basic-level categories (Rosch, 1975; Rosch, Mervis, et al., 1976).

Superordinate categories experiments

The first experiment consisted of listing attributes possessed by each exemplar from the following superordinate categories: "furniture", "vehicle", "fruit", "weapon", "vegetable", and "clothing". Two judges evaluated collected attributes (e.g., fill attribute to other categories when was mentioned only for one). The main goal was to assess the first part of the family resemblance hypothesis.

Each attribute received a weight according to its presence in category exemplars (e.g., if ten from twenty exemplars possess the attribute, it gets a score of 10). Then, they calculated the family resemblance score of the exemplar by summing the weights of individual attributes the exemplar possesses (e.g., an exemplar with three attributes, each with a weight of 3, would have the family resemblance score equal to 9). They also proposed a second variation of the family resemblance measure, which included a natural logarithm of attribute weights. This modification was motivated by the possibility that the difference between an attribute possessed by two exemplars versus one exemplar is not equal to the difference between an attribute possessed by 19 versus 18 exemplars.

They calculated Spearman rank-order correlations between the calculated family resemblance scores and previously collected typicality ratings[1] to test their hypothesis. The correlations showed a strong relationship between family resemblance and human typicality ratings:

> These correlations, for the basic measure of family resemblance, were: furniture, 0.88; vehicle, 0.92; weapon, 0.94; fruit, 0.85; vegetable, 0.84; clothing, 0.91. These correlations for the logarithmic measure of family resemblance were: furniture, 0.84; vehicle, 0.90; weapon, 0.93; fruit, 0.88; vegetable, 0.86; clothing, 0.88. All were significant ($p < .001$). (Rosch and Mervis, 1975)

Later studies suggested that these correlations are overestimated (Barsalou, 1987). Nevertheless, their importance is still relevant.

The second experiment aimed to verify the later part of the family resemblance hypothesis. Firstly, they had to collect the so-called *contrast categories*. The contrast categories are on the same hierarchical level but have contrasting linguistic meanings. The empirical method in which participants answered questions in the form of "If X is not a Y, what is it (might be)?" was used (Frake, 2012; Rosch and Mervis, 1975). Rosch and Mervis found that the contrast categories obtained were too inconsistent across participants. Therefore, they had to test the second half of the family resemblance indirectly:

> If the best examples of superordinate categories are those with least in common with other categories they should be dominant members of few (or no) categories other than the superordinate in question. Thus, prototypicality should be correlated with a measure of the dominance of a category over its members.... The hypothesis of Experiment 2 was, thus, that the more prototypical a member of a superordinate category, the less dominant its membership would prove to be in categories other than the superordinate in question.(Frake, 2012; Rosch and Mervis, 1975)

The respondents completed the questionnaire, writing three categories to which a given noun (member from one of the superordinate categories) belongs. Rosch and Mervis then gave each category weight according to the order in which respondents listed them. The category dominance for each item was calculated the following way: The designated superordinate minus the most frequently mentioned other superordinate, plus the designated superordinate minus the second most frequently measured other superordinate (Rosch, 1975).

The Spearman rank-order correlations of item dominance and typicality scores were calculated (note, the category clothing was erroneously omitted):

> These correlations were: fruit, 0.71; furniture, 0.83; vegetable, 0.67; vehicle, 0.82; weapon, 0.77. All were significant ($p < .001$). (Rosch, 1975)

[1]Typicality ratings were collected from a selected group of respondents and judged on the 7-point scale (Rosch, 1975).

Therefore, they concluded that the second part of the family resemblance hypothesis is confirmed.

Basic level categories experiments

The second part of the paper examines the family resemblance hypothesis on the basic level categories. The following basic-level categories were selected: "car", "truck", "airplane", "chair", "table", and "lamp". Since basic-level categories are objects with a possible infinite population (Rosch and Mervis, 1975), 15 pictures represented each category.

The participants did attribute listing in a similar way as in the first experiment. The main difference was that the pictures were used as stimuli instead of words. The goodness of example ratings were gathered on a 7-point scale, similarly to superordinate categories (Rosch, 1975).

The typicality ratings were calculated in a similar way as in the first experiments. As expected, the main difference was the larger number of shared attributes across members of basic-level categories. The Spearman rank-order correlations showed the strong relationship between the family resemblance measures and goodness of example:

> ...the basic measure of family resemblance and prototypicality were: car, 0.94; truck, 0.84; airplane, 0.88; chair, 0.81; table, 0.88; and lamp, 0.69. The correlations between the logarithmic measure of family resemblance and prototypicality were: car, 0.86; truck, 0.88; airplane, 0.88; chair, 0.79; table, 0.85; and lamp, 0.64. All were significant ($p < .01$).

Like the second experiment, the fourth experiment examined the second part of the family resemblance hypothesis. The main difference was that the hypothesis's second part was directly verifiable for basic-level categories. The experiment was separated into three parts: determining which categories are the contrasting ones, obtaining a list of attributes for pictures from contrasting categories, and calculating the correlation between the number of attributes shared with items from contrasting categories with gathered typicality ratings.

The contrast categories were obtained by asking questions like "If X is not a Y, what is it (might be)?". The attribute listing was done in a similar way as in the third experiment. The Spearman rank-order correlations were calculated between typicality ratings and attribute overlap ranks:

> A Spearman rank-order correlation was performed between the prototypicality and attribute overlap ranks of the 15 chair and 15 car pictures. Results were: chairs $r = -0.67$; cars, $r = -0.86$. Both were significant ($p < .01$).

Artificial categories experiments

Since the everyday experience of respondents influences the natural categories used in the previous two parts, two experiments with artificial categories were conducted. These categories were artificially constructed to follow the family resemblance hypothesis: items from the categories differed only in the degree of family resemblance within categories or size of overlap of attributes between categories (Rosch and Mervis, 1975). The hypothesis was that this structure of categories affects the learning rate, reaction time in categorizing, and goodness of example ratings.

The fifth experiment evaluates the first part of the family resemblance hypothesis in the context of artificial categories. The strings of letters[2] were constructed as stimuli (e.g., HPNWD, HPNSJ). The family resemblance score was calculated similarly to exemplars with attributes. Two groups of artifact categories with six members were prepared, one with a symmetric structure and one with an asymmetric structure. The symmetric structure includes two central strings with equal family resemblance, two central, intermediate, and two extreme exemplars according to family resemblance scores. For the asymmetric structure, all members of categories have different family resemblance scores. Rosch and Mervis found that items with greater family resemblance are learned, identified, and judged faster than those with less resemblance.

The sixth experiment examined the second part of the family resemblance hypothesis. Only the symmetrically structured group of categories was used because the asymmetric one could not learn within one hour of subject time. They have shown that the extent of overlap with contrast categories influences category structure for both items, which differ and do not differ in terms of inner family resemblance.

2.2.2 Central tendency, frequency, and ideals

Barsalou (1985) provided essential insights into natural and goal-derived categories. He examined the typicality in common taxonomic categories and goal-derived categories, which were especially important since Hampton (1981) found that family resemblance did not predict typicality well in the case of some abstract categories. Barsalou proposed new determinants of typicality: *ideal* (how well the exemplar serves a category goal), *central tendency* (how similar is the exemplar to the category prototype), and *frequency of instatiation* (subjective estimates of how often is the exemplar experienced as a category member).

Central tendency

The central tendency can be understood as central information about the category (e.g., the average or median of the category exemplars). This provides another possible view on family resemblance proposed by Rosch and Mervis (1975). Instead of defining it as the overlap of attributes, Barsalou (1985) described it as an exemplar's similarity to the central tendency because average similarity to other category members must be roughly the same (Barsalou, 1983). As Murphy noted (G. Murphy, 2004, p. 35), the "family resemblance" sometimes refers only to the first part of the original hypothesis from Rosch and Mervis – that was the case with Barsalou's proposition.

Rosch and Mervis (1975) did not explicitly define family resemblance in terms of similarity or distance; the term "overlapping attributes" was used throughout the paper. Nevertheless, they mentioned the possibility of interpreting this as a similarity of the exemplars:

[2]Digits were used only when more symbols were needed (Rosch and Mervis, 1975)

> If, in addition, items are perceived as similar to each other in proportion to the number of attributes which they have in common, multidimensional scaling of the similarity judgments between all pairs of items in a category should result in a semantic space in which the distance of items from the origin of the space is determined by their degree of family resemblance. (Rosch and Mervis, 1975)

The main argument for this definition was effectivity since comparing the exemplar to one central tendency is more effective than comparing it to all other category members.

For two reasons, the central tendency is expected to be a more vital determinant for the common taxonomic categories than goal-derived categories. Firstly, common taxonomic categories possess so-called *correlational structure* (Barsalou, 1985). Attributes from common taxonomic categories usually co-exist with other properties, which is usually not true for goal-oriented categories. Secondly, common taxonomic categories are generally used for categorization. Therefore, they are optimized for classification performance.

Ideals

Barsalou (1985) proposed the concept of ideals to account for how well a given exemplar serves the goals of the category. He illustrated his hypothesis with an example of a category "food which to eat on a diet", which is expected to have an ideal "zero/low calories" (Barsalou, 1985). According to Barsalou, a single category can have more than one ideal; most of them are expected to have more than one.

Barsalou emphasized the context-dependence of ideals. He proposed that ideals are expected to change when the goal changes. For example, the piano is viewed differently when our goal is to move it to the fourth floor (it is heavy) versus performing a piece at the concert (sound quality) (Barclay et al., 1974).

Frequency of Instantiation

The frequency of instantiation can be defined as a subjective estimate of how often someone has experienced an entity as a category member (Barsalou, 1985). Barsalou compared it to *familiarity*, which can be defined as the estimation of how often an exemplar has been experienced across all contexts. The main difference is that frequency of instantiation is a category-specific measure of frequency, while familiarity is a category-independent measure of frequency.

Predictive power experiment

The first experiment examined if central tendency, ideals, and frequency of instantiation predict typicality in goal-derived and common taxonomic categories.

Nine goal-derived categories were used: "birthday presents", "camping equipment", "transportation for getting from San Francisco to New York", "personality characteristics in people that prevent someone from being friends with them", "things to do for weekend entertainment", "foods not to eat on a diet", "clothes to wear in the snow", "picnic activities", and "things to take from one's home during a fire".

Nine common taxonomic categories were used (same as in Rosch and Mervis (1975) study): “vehicles”, “clothing”, “birds”, “weapons”, “vegetables”, “sports”, “fruit”, “furniture”, and “tools”.

First, the exemplars for all of the 18 categories were gathered. Then, for all exemplars, the goodness-of-example (which can be understood as typicality), frequency of instantiation, ideals, and family resemblance was gathered. The goodness-of-example was rated on a 9-point scale ranging from “poor example” to “excellent example”. The frequency of instantiation used a similar 9-point scale ranging from “not frequently at all” to “very frequently”. The ideals were rated on a specific 9-point scale per category ranging from “very low amount” to “very high amount” (e.g., “how happy people are to receive it” in the case of “birthday presents”). Lastly, the family resemblance score was calculated for each exemplar by averaging the similarity across all possible pairs of exemplars within the given category, rated on a 9-point scale ranging from “not similar at all” to “very similar”.

The multiple correlations were computed across categories. The average correlation between ideals and goodness of exemplar was stronger for goal-derived categories (0.70) than for common taxonomic categories (0.46). Similarly, the average correlation between the frequency of instantiation and goodness of exemplar was more significant for goal-derived categories (0.56) than for common taxonomic ones (0.49). As predicted, the average correlations between central tendency and goodness of exemplar were stronger for taxonomic categories (0.63) than for goal-derived categories (0.38).

The partial correlations were calculated to examine the possibility that previous correlations were caused by variance shared between three possible predictors: central tendency, ideals, and frequency of instantiation. In other words, a high correlation of one of the predictors can likely be caused by shared variance with some other strongly correlated predictor. For goal-derived categories, the relation between the goodness of the example and the central tendency was almost eliminated (correlation dropped to 0.05). The situation of common taxonomic categories was the opposite; central tendency became an even stronger predictor of the goodness of an example. However, the ideals and frequency of instantiation also accounted significantly for the unique goodness of example variance. They were similar for goal-derived categories and common taxonomic ones.

Context dependence experiment

The second experiment examined if ideals determine typicality and if any of the three determinants depend on the context. Murphy described it as a problem of not knowing which is the chicken and which is the egg since ideals may arise from typical exemplars of the category instead of being the factor in forming the category (G. Murphy, 2004, p. 37).

Participants learned two artificial categories. Members of these categories were the person’s last names (e.g., Davis, Wilson) with corresponding spare time activities (e.g., dance, renovate houses, write poetry, go to movies). The two categories were designated alongside two variables: “jogging” and “reading the newspaper”. All members of one category jogged, and all members of the second category spent their free time reading the newspaper. Exemplars varied in two aspects. Firstly, they differed in the frequency of spare time activities (daily, weekly, monthly). Secondly, they varied in similarity to their central tendency, having a high, medium, or low number of characteristic activities.

One group of participants (related dimension) was told that the category formed by "jogging" people belong to the category of "physical education teachers" and that the category formed around "reading the newspaper" belongs to "current event teachers". Participants were expected to follow the ideals that physical education teachers are usually physically active and that current event teachers spend their time reading about the news.

The second group of participants (unrelated dimension) was told that the category defined by "jogging" are "Q language programmers" and the second category defined by "reading the newspaper" are "Z language programmers". People are expected only to know ideals related to Q or Z programming language. Thus, ideal should not determine typicality.

Barsalou found that ideals and central tendency affect typicality. However, each determinant was affected by dimension. Ideals have a significant effect in related dimension cases, but no effect was found in the case of unrelated dimensions. The effect of central tendency was the opposite. An important effect was found in the case of an unrelated dimension, and a minor effect was found in the case of a related dimension.

2.2.3 Structural and functional typicality

With an increasing number of neuroscience research, Dieciuc and Folstein (2019) reviewed the results from typicality effects research and proposed a framework of *structural typicality* and *functional typicality*. They argued that a stable approach to typicality presented by Rosch and Mervis (1975) can coexist with the more dynamic way of ideals presented by Barsalou (1985).

Structural typicality

Dieciuc and Folstein (2019) described the structural typicality as the result of encoding relative consistent correlations between features of exemplars from our world into long-term memory (Dieciuc and Folstein, 2019).

The stable structure of typicality found strong support in the controlled semantic cognition (Section 1.2.7) and semantic dementia research. Mayberry, Sage, and Ralph (2011) found that patients with semantic dementia make typicality-related errors during a categorization task. Atypical exemplars were not categorized (e.g., the penguin was not a bird), and "pseudotypical" nonmember exemplars were incorrectly categorized (e.g., the butterfly was a bird) (Dieciuc and Folstein, 2019; Mayberry, Sage, and Ralph, 2011). Identical results were replicated in drawing task (Bozeat et al., 2003), where participants omitted distinctive features of atypical exemplars and overextended features of typical ones.

Similar results were reproduced on healthy individuals during a picture-naming task (Woollams, 2012). Participants had temporarily inhibited brain regions by the TMS, which is usually affected by semantic dementia, which made them slower in picture-naming tasks of atypical exemplars.

The hypothesis which assumes that category structure emerges from the correlation of features is supported by fMRI research. Iordan et al. (2016) found direct support for the family resemblance approach to typicality by examining the intra-category neural structure of eight common taxonomic categories. They found that neural activity elicited by more typical exemplars is more similar to the central tendency (first part of the family resemblance), and neural activity produced by atypical exemplars is more similar to the central tendency of other categories (second part of the family resemblance). Moreover, they found brain regions in which the neural representation space was organized according to the atypical exemplars. This finding was expected because even atypical exemplars are still categorized as members of the category, and the ability to distinguish them may arise from this region.

These results are significant since, despite the influence of the typicality effect, only behavioral studies were conducted before. They concluded that these findings may support that the brain uses both exemplars (Section 1.2.3) and prototype theories (Section 1.2.2) to organize categories.

Functional typicality

The functional typicality emerges from observing how a subset of information is retrieved from the long-term to the working memory in a task-dependent and context-specific manner (Dieciuc and Folstein, 2019). As a result, the typicality could be observed as unstable and dynamic.

The previously discussed ideals introduced by Barsalou play an essential role in the instability of typicality across cultures. Ojalehto and Medin (2015) found that culture strongly influences ideals, and Dieciuc and Folstein (2019) stated that culture is inseparable from the conceptual structure of categories.

Expertise is another factor that influences the stability of the typicality. Most of the previously mentioned studies worked with undergraduate students as subjects (Barsalou, 1983, 1985; De Deyne et al., 2008; Hampton, 1979, 1981, 1982; Mervis and Pani, 1980; Rosch, 1975; Rosch and Mervis, 1975; Rosch, Mervis, et al., 1976). Bailenson et al. (2002), Burnett et al. (2005), E. B. Lynch, Coley, and Medin (2000), and Medin and Atran (2004) conducted studies with experts and found that long-term experience shapes the typicality and makes ideals a much more vital determinant of typicality. On the other hand, undergraduate or novice subjects almost wholly rely on familiarity. These results suggest that typicality is influenced by knowledge about the exemplars obtained during the learning process instead of family resemblance (Dieciuc and Folstein, 2019).

The ad-hoc categories that are not directly stored inside long-term memory are formed “on the fly” according to the current goal (Dieciuc and Folstein, 2019). Therefore typicality changes much faster than the taxonomy categories, which are organized according to the much more stable family resemblance.

Freeman (2014) showed that the typicality could change across different contexts. Roth and Shoben (1983) and Yeh and Barsalou (2006) found that agreement between respondents can be increased by setting specific context during the typicality gathering tasks. This raises concerns in the case of studies that did not specify the context (e.g., Rosch (1975), or Barsalou (1985)). Subjects may approach the task completely contextless, but on the other hand, they could be influenced by some context they came with. In the former case, the stability of the typicality can be an artifact that arises from averaging the subject’s responses (Dieciuc and Folstein, 2019).

The perspective is another cause of instability of typicality (Barsalou and Sewell, 1984; Hampton, Dubois, and Yeh, 2006; Kim and G. L. Murphy, 2011). Barsalou and Sewell

(1984) found that typicality changes across the subjects' perspectives. For example, subjects agreed that vegetarian food is a typical food from "hippies" point of view (Dieciuc and Folstein, 2019). This taught an important lesson: the subjects can consciously manipulate typicality by taking a different point of view.

Working together

An important conclusion of Dieciuc and Folstein's work is that structural and functional typicality are not mutually exclusive:

> Structural typicality reflects organization of information in semantic memory based on similarity. Functional typicality reflects the ability to construct, deconstruct, and reconstruct our concepts in multiple ways depending upon the current needs of the observer and the current affordances of the environment. (Dieciuc and Folstein, 2019)

They demonstrated this in the following analogy:

> There is a fairly systematic order to the way I arrange dishes in my kitchen cabinets. Small and large plates go together, cups and mugs go together, but how I use my dishes depends on what meal I am having. Some meals require large plates, some require small plates plus a bowl. The structure by which my dishes are arranged in the cabinet is fairly stable. Conversely, the function of how I selectively recruit the dishes is fairly flexible (unstable). Big plates and small plates might go together in the cabinet (structurally), whereas small plates and bowls might go together when eating (functionally). In this same way, two things may be similar in structural typicality while simultaneously being dissimilar in functional typicality (or vice versa). If I was told to set the table for dinner but I did not know what kind of meal I was having, I might grab the plates and utensils that are most easily accessible. But if I knew we were having something specific, I would then reach for the dishes that were most relevant, even if they were less accessible. (Dieciuc and Folstein, 2019)

This is not an entirely new idea. As we described earlier, Barsalou (1985) found that central tendency and ideals are not mutually exclusive determinants of typicality. Nevertheless, it is important to reflect and examine the typicality phenomenon in the light of new theories like situated theory (which is omitting the central core of concepts altogether, see Section 1.2.6) and controlled semantic cognition (which brings evidence of the existence of stable concept core; see Section 1.2.7).

An important implication of this framework is that it can unify research that could be understood as contradictory otherwise (Dieciuc and Folstein, 2019). According to Dieciuc and Folstein, more research is needed to examine how structural and functional typicality work together, mainly to determine which is more dominant in the given moment.

2.3 Conclusions

We have shown that the typicality effect is among the strongest and most reliable effects across multiple fields. Its influence can be seen in category judgment, category learning, category inference, exemplar generation frequency, linguistics, and many more.

Significant results were obtained in behavioral research of possible determinants. Rosch and Mervis proposed the family resemblance hypothesis, which changed the approach to understanding the structure of categories. Barsalou demonstrated that family resemblance is not the only determinant of typicality and that typicality can be highly dynamic depending on the context. More recently, Dieciuc and Folstein proposed the structural and functional typicality framework supported by neuroimaging and behavioral research, which provides a unified ground for examining stable and dynamic effects of the typicality.

Chapter 3

Formalization of typicality

In this chapter, we will summarize the approaches to formalization of typicality proposed by Belohlavek and Mikula (2022, 2024a,b,c,d). These papers are available as appendices A–E. Our formalizations are rooted in the psychology of concepts. We believe that the data science field can strongly benefit from the findings available in the field of psychology. On the other hand, psychology can benefit from multidisciplinary research, which can bring back some interesting insights.

3.1 Framework preliminaries

Before we dive into the formalizations of typicality, we will present the essential preliminaries relevant to our research on typicality.

The core component is the formal concept analysis (FCA) framework (Belohlavek, 2008; Carpineto and Romano, 2004; Ganter and Wille, 2012; Wille, 1982). The FCA is naturally connected to cognitive psychology since, as we will see, it is firmly rooted in traditional logic and classical theory (Belohlavek, 2008). Based on our discussion in Chapter 1, it would be naive to understand it as a precise model of the human conceptual system. Nevertheless, it can still provide a simple and robust framework for formalizing basic concepts from the psychology of concepts.[1]

3.1.1 Input data

Input data forms a binary matrix, representing the relation between objects and their attributes. Formally, it is called *formal context*.

Definition 1. *A formal context is a triplet $\langle X, Y, I \rangle$ where X is a non-empty set of objects, Y is a non-empty set of attributes, and I is a binary relation between X and Y, i.e., $I \subseteq X \times Y$. The $\langle x, y \rangle \in I$ represents that an object $x \in X$ has an attribute $y \in Y$.*

A formal context represents our universe; from this point of view, no more objects and attributes exist. This is essential to recognize since the quality of our results and, ultimately, the formalizations and their evaluation lie in these input data.

[1]The framework was previously used for the formalization of basic level phenomenon (Belohlavek and Trnecka, 2020a,b).

3.1.2 Formal concepts

Formal concepts are formed inside our formal universe, represented by a formal context. A *formal concept* will denote a concept in the sense of classical theory – necessity and sufficiency – an object is part of a formal concept when it possesses all defining attributes and at the same time, definitional attributes are jointly sufficient to an object is a member of a formal concept.

Definition 2. *Given formal context $\langle X, Y, I\rangle$ a pair $\langle A, B\rangle$ where $A \subseteq X$ and $B \subseteq Y$ is called a formal concept in $\langle X, Y, I\rangle$ if and only if $A^{\uparrow} = B$ and $B^{\downarrow} = A$ where*

$$\begin{aligned} A^{\uparrow} &= \{y \in Y \mid \textit{ for each } x \in X : \langle x, y\rangle \in I\}, \\ B^{\downarrow} &= \{x \in X \mid \textit{ for each } y \in Y : \langle x, y\rangle \in I\}. \end{aligned}$$

For a formal concept $\langle A, B\rangle$ a set of all objects $A = B^{\downarrow}$ within a formal concept is called a *extent*, and the set of all attributes $B = A^{\uparrow}$ is called a *intent*.

3.1.3 Conceptual hierarchy

Formal concepts can be ordered into a *conceptual hierarchy* (Ganter and Wille, 2012).

Definition 3. *For formal concepts $\langle A_1, B_2\rangle$ and $\langle A_2, B_2\rangle$ of $\langle X, Y, I\rangle$, put $\langle A_1, B_2\rangle \leq \langle A_2, B_2\rangle$ if and only if $A_1 \subseteq A_2$ (or dualy iff $B_1 \supseteq B_2$).*

Concept ordering captures the intuition that the concept “dog” is *subconcept* of concept “mammal” and the other way the concept “mammal” is *superconcept* of the concept “dog” (Belohlavek, 2008). Conceptual ordering is another parallel to classical theory where logically defined concepts allow a simple way of concept hierarchies with transitivity (see Section 1.2.1).

The collection of all formal concepts of the given formal context is called *concept lattice* (Belohlavek, 2008).

Definition 4. *Denote by $\mathcal{B}(X, Y, I)$ the collection of all formal concepts of $\langle X, Y, I\rangle$, ie.*

$$\mathcal{B}(X, Y, I) = \{\langle A, B\rangle \in 2^X \times 2^Y | A^{\uparrow} = B, B^{\downarrow} = A\}$$

A $\mathcal{B}(X, Y, I)$ equipped with concept ordering $\leq$ denoted as $\langle \mathcal{B}(X, Y, I), \leq\rangle$ is called concept lattice of $\langle X, Y, I\rangle$.

Concept lattice can be loosely understood as an analogy to human conceptual hierarchy. Note that concept lattice includes all logically plausible concepts with respect to formal context, even those that seem unreasonable to us.

3.1.4 Similarity

The following general definition of similarity covers the wide range of *similarity measures.*

Definition 5. *Similarity measure on a set X of objects is a binary function*

$$sim : X \times X \to \mathbb{R} \tag{3.1}$$

The value $sim(x, y)$ is interpreted as the extent to which x is similar to y.

Two schemes usually define the similarity measures described in the literature. Let us illustrate that with an example of a well-known Jaccard index (Jac) coefficient.

The first way of defining similarity assumes that objects are described with subsets of the attributes from the attribute universum Y. Therefore, we define our similarity between object x and y as

$$sim(x, y) = sim_Y(B_1, B_2), \tag{3.2}$$

where

$$sim_Y : 2^Y \times 2^Y \to \mathbb{R}$$

and $B_1, B_2 \subseteq 2^Y$ are two sets of attributes representing objects x and y. According to this scheme, the Jaccard index can be defined as

$$sim_{\mathrm{Jac}}(B_1, B_2) = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|}. \tag{3.3}$$

The second way assumes the set X of all objects described by n binary attributes identified with the set $\{0, 1\}^n$ of all n-dimensional binary vectors. Let us have two binary vectors $x, y \in \{0, 1\}^n$; any similarity can be then defined with the use of the following scheme:

$$\begin{array}{c|cc|c} & y=1 & y=0 & \Sigma \\ \hline x=1 & a & b & a+b \\ x=0 & c & d & c+d \\ \hline \Sigma & a+c & b+d & n \end{array} \tag{3.4}$$

in which, e.g., a is the number of attributes i for which $x_i = 1$ and $y_i = 1$, b is the number of i for which $x_i = 1$ and $y_i = 0$, $a + c$ is the number of attributes for which $y_i = 0$, etc.

Following this scheme, the Jaccard index can be defined by a formula involving a, b, c, and d, which correspond to $x, y \in \{0, 1\}^n$ according to (3.4)

$$sim_{\mathrm{Jac}}(x, y) = \frac{a}{a + b + c}. \tag{3.5}$$

3.2 Defining typicality

Our approach to typicality starts with examining the family resemblance hypothesis proposed by Rosch and Mervis (1975). We propose multiple schemes to the typicality and study their mutual relationship. The following sections will provide an overview of our results.

3.2.1 Rosch and Mervis' scheme

Rosch and Mervis (1975) verbally described their way of calculating the typicality in their experiments as follows:

> ... each attribute was weighted in accordance with the number of items in the category possessing it. The basic measure of the degree of family resemblance for an item was the sum of the weighted scores of each of the attributes that had been listed for that item. (Rosch and Mervis, 1975)

Note that this description covers only the first part of the family resemblance (see Section 2.2.1) and does not take the outside of the category into account. The second part of this hypothesis was tested indirectly in the case of basic-level categories and in the form of attribute overlap in the case of superordinate categories.[2]

The core part of their verbal description is the weighted score of attributes. We can naturally define *weight of the attribute* for a nonempty set $A \subseteq X$ representing a category and an arbitrary attribute $y \in Y$ as follows (Belohlavek and Mikula, 2022, 2024c):

$$w(y, A) = |\{x\,;\, x \text{ is in } A \text{ and has } y\}|. \tag{3.6}$$

Note that this general definition does not put any requirements on A being the extent of a formal concept $\langle A, B\rangle$. Thus, one can rewrite the definition for formal concept $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$ as follows:

$$w(y, \langle A, B\rangle) = |\{x \in A \mid x \in \{y\}^{\downarrow}\}|. \tag{3.7}$$

The weight of the attribute allows us to define the degree of typicality as a sum of all weights of its attributes (Belohlavek and Mikula, 2022, 2024c):

Definition 6. *The degree $typ_{RM}(x, A)$ of typicality of an object x in A is then defined by*

$$typ_{RM}(x, A) = \sum_{y \in Y,\ \langle x,y\rangle \in I} w(y, A), \tag{3.8}$$

i.e., as the sum of weights of all the attributes y possessed by x.

One can again rewrite the definition for formal concept $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$ simply as:

$$typ_{RM}(x, \langle A, B\rangle) = \sum_{y \in \{x\}^{\uparrow}} w(y, \langle A, B\rangle). \tag{3.9}$$

3.2.2 Extending Rosch and Mervis' formula

Rosch and Mervis' formula presented in the previous section can be naturally extended into the more general formula. This extension is based on the fact that the original

[2]Because all methods mentioned above require additional information (e.g. the list of contrast categories) and previous research found that the first part of the family resemblance is more influential (G. Murphy, 2004), we take this verbal description as the main source of our formalization.

formula counts only the presence of attributes. We thus propose generalization, which takes the absence of the attributes in the account (Belohlavek and Mikula, 2024c).

The Rosch and Mervis' weight (3.6), now denotes $w^+(y, A)$ is accompanied by the symmetric weight $w^-(y, A)$, which represents the number of objects A that do not have the attribute y:

$$\begin{aligned} w^+(y, A) &= |\{x\,;\, x \text{ is in } A \text{ and has } y\}|, \text{ and} \\ w^-(y, A) &= |\{x\,;\, x \text{ is in } A \text{ and does not have } y\}|. \end{aligned}$$

Since people usually regard attribute presence as more significant than absence, we proposed using non-negative weights a^+ and a^- that allow us to set the significance of shared presences and shared absences differently. This allows us to define the generalized scheme of Rosch and Mervis' formula:

Definition 7. *For the non-negative weights a^+ and a^-, the degree $typ_{RM^\pm}^{a^+,a^-}(x, A)$ of typicality of an object x in A is then defined by*

$$typ_{RM^\pm}^{a^+,a^-}(x, A) = a^+ \cdot \sum_{y \in Y,\ \langle x,y\rangle \in I} w^+(y, A) + a^- \cdot \sum_{y \in Y,\ \langle x,y\rangle \notin I} w^-(y, A). \tag{3.10}$$

For $a^+ = 1$ and $a^- = 0$, the formula (3.10) yields the original Rosch and Mervis' typicality formula.

3.2.3 Similarity scheme

Let us now revisit the first part of the family resemblance hypothesis proposed by Rosch and Mervis (1975):

> ... members of a category come to be viewed as prototypical of the category as a whole in proportion to the extent to which they bear a family resemblance to (have attributes which overlap those of) other members of the category. (Rosch and Mervis, 1975)

Rosch and Mervis did not explicitly define family resemblance in terms of similarity or distance; the term "overlapping attributes" was used throughout the paper. Nevertheless, they mentioned the possibility of interpreting this as a similarity of the exemplars:

> If, in addition, items are perceived as similar to each other in proportion to the number of attributes which they have in common, multidimensional scaling of the similarity judgments between all pairs of items in a category should result in a semantic space in which the distance of items from the origin of the space is determined by their degree of family resemblance. (Rosch and Mervis, 1975)

Thus, the most typical object from the given category is the one that is the most similar to the other objects from the category. Also, let us recall the central tendency hypothesis from Section 2.2.2, which was also defined in terms of similarity. Defining typicality this way allows us to use a more general scheme that can utilize various similarity measures.

Therefore, let us define the *degree of typicality* as follows (Belohlavek and Mikula, 2022, 2024c):

Definition 8. *Given a similarity* $sim : X \times X \to \mathbb{R}$*, an object* $x \in X$*, and a nonempty set* $A \subseteq X$ *representing a category, a degree of typicality of* x *in* A *is defined by*

$$typ(x, A) = \frac{\sum_{x_1 \in A} sim(x, x_1)}{|A|}. \tag{3.11}$$

Note that this definition is directly applicable to the formal concepts since A can be understood as the extent of formal concepts $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$.

3.2.4 Relationship between definitions

The original Rosch and Mervis' formula (Definition 6), extended version of Rosch and Mervis' scheme (Definition 7), and similarity-based scheme (Definition 8) are on the first sight quite different approaches to the typicality formalization. Yet, the following theorem and corollaries show that Rosch and Mervis' formula and its extension are both the result of a particular scaling of similarity-based scheme with appropriately selected similarity coefficients (Belohlavek and Mikula, 2022, 2024c).

For our purpose, we use the notation from formal concept analysis and denote by $\{x\}^{\uparrow}$ the set of all attributes shared by the object x, i.e.,

$$\{x\}^{\uparrow} = \{y \,;\, x \text{ has } y\}.^{3}$$

We consider the similarity function

$$\mathrm{SMC}^{a^+,a^-}(x_1, x_2) = \frac{a^+ \cdot |\{x_1\}^{\uparrow} \cap \{x_2\}^{\uparrow}| + a^- \cdot |Y - (\{x_1\}^{\uparrow} \cup \{x_2\}^{\uparrow})|}{|Y|} \tag{3.12}$$

parameterized by non-negative weights a^+ and a^-.

Theorem 1. *For arbitrary* $a^+, a^- \geq 0$*, each object* x*, and any category* A*,*

$$typ^{a^+,a^-}_{RM^{\pm}}(x, A) = |A| \cdot |Y| \cdot typ_{\mathrm{SMC}^{a^+,a^-}}(x, A)$$

where $typ_{\mathrm{SMC}^{a^+,a^-}}(x, A)$ *is determined by* SMC^{a^+,a^-} *according to (3.11).*

Let us now discuss two important corollaries of the Theorem 1 (Belohlavek and Mikula, 2024c). For this purpose, we consider two particular choices of a^+ and a^-:

(a) $a^+ = 1$ and $a^- = 0$: In this case, the similarity function in (3.12) shall be denoted RR, i.e.,

$$\mathrm{RR}(x_1, x_2) = \mathrm{SMC}^{1,0}(x_1, x_2) = \frac{|\{x_1\}^{\uparrow} \cap \{x_2\}^{\uparrow}|}{|Y|}.$$

[3]Note that the following theorems do not require categories to be formal concepts; we use this notation for convenience.

The function RR is, in fact, one of the existing similarity measures, called the Russel-Rao coefficient (Belohlavek and Mikula, 2024b).

(b) $a^+ = 1$ and $a^- = 1$: In this case, the similarity in (3.12) shall be denoted SMC, i.e.,

$$\mathrm{SMC}(x_1, x_2) = \mathrm{SMC}^{1,1}(x_1, x_2) = \frac{\cdot|\{x_1\}^\uparrow \cap \{x_2\}^\uparrow| + |Y - (\{x_1\}^\uparrow \cup \{x_2\}^\uparrow)|}{|Y|}.$$

This function is one of the best-known similarity measures known as the simple matching coefficient (SMC) or the Sokal-Michener coefficient (Belohlavek and Mikula, 2024b).

The following corollaries of Theorem 1 show that the long-established similarity measures RR and SMC are precisely the measures corresponding to the original Rosch and Mervis' formula typ_{RM} and its presence/absence extension $typ_{\mathrm{RM}^\pm}$.

The first corollary relates to Rosch and Mervis' formula:

Corollary 1. *For each object x and an arbitrary category A,*

$$typ_{\mathrm{RM}}(x, A) = |A| \cdot |Y| \cdot typ_{\mathrm{RR}}(x, A)$$

where typ_{RR} is the typicality (3.11) induced by the Russell-Rao coefficient.

The second corollary relates to the extension of Rosch and Mervis' formula:

Corollary 2. *For each object x and an arbitrary category A,*

$$typ_{RM^\pm}(x, A) = |A| \cdot |Y| \cdot typ_{\mathrm{SMC}}(x, A)$$

where typ_{SMC} is the typicality (3.11) induced by the simple matching coefficient.

The above theorem and its corollaries provide better insights into the original Rosch and Mervis' formula by confirming its equivalence to the similarity-based scheme. Corollary 1 shows that Rosch and Mervis' scheme encodes the well-known Russel-Rao similarity coefficient without explicitly mentioning it. This is a remarkable result for the psychology of concept and the Russel-Rao similarity, which received psychological support for its significance. Corollary 2 shows a relationship between the similarity-based scheme with the simple matching coefficient and an extended version of Rosch and Mervis' scheme.

In the next section, we will show that even the Rosch and Mervis' scheme can be extended to account for the information outside the object's category.

3.2.5 Attribute characteristicness as a weight

The basic and extended Rosch and Mervis' formulas involve attribute weights, which utilize only the information within the category. All previously defined weights $w(y, A)$, $w^+(y, A)$, and $w^-(y, A)$ utilize only the information from the objects to which y applies within the category A. This limitation did not restrict the original theory since Rosch and Mervis proposed this weight to capture the attribute's presence within the category. On the other hand, Rosch and Mervis mentioned information outside of the category in their family resemblance hypothesis:

> ... Conversely, items viewed as most prototypical of one category will be those with (2) least family resemblance to or membership in other categories. (Rosch and Mervis, 1975) (added parts numbering)

The second part was omitted entirely from their original formula (3.7) and was tested only in the form of attribute overlap with contrast categories in the case of basic-level categories and as category dominance in the case of subordinate categories. Both approaches require more information about the category and thus cannot be applied to objects with only attribute data.

Hence, we propose a new attribute weight that utilizes information outside the category and does not require additional information about the categories. The form of attribute weight derives from a natural concept of *characteristicness of an attribute*. Similar ideas have been described in the literature, although mainly in a verbal manner, under various names, including distinctiveness (Bozeat et al., 2003; Garrard et al., 2001), centrality (W.-k. Ahn et al., 2000; Sloman and W.-K. Ahn, 1999; Sloman, Love, and W.-k. Ahn, 1998; Ward et al., 2000), diagnosticity (Hsu, Schlichting, and Thompson-Schill, 2014; Chin-Parker and Ross, 2004), and attribute typicality (Woollams, 2012).

Put briefly, we consider an attribute y characteristic of a category A to the extent to which a member of A is roughly equivalent to having y (Belohlavek and Mikula, 2024c). For our purpose, we denote by $\{y\}^{\downarrow}$ the set of all objects shared by the attribute y, i.e.,

$$\{y\}^{\downarrow} = \{x\,;\, x \text{ has } y\}.$$

This idea offers two possible ways to formalize it. The first way leads to the formula

$$w(y, A) = \frac{|\{y\}^{\downarrow} \cap A|}{|A|} \cdot \frac{|\{y\}^{\downarrow} \cap A|}{|\{y\}^{\downarrow}|}. \tag{3.13}$$

The first factor, $\frac{|\{y\}^{\downarrow} \cap A|}{|A|}$, can be interpreted as the first part of the family resemblance – the extent to which the objects in A have attribute y. The second factor, $\frac{|\{y\}^{\downarrow} \cap A|}{|\{y\}^{\downarrow}|}$, captures the essence of the second part of the family resemblance hypothesis and may be regarded as the extent to which the objects sharing y belong to category A.

The second way leads to

$$w(y, A) = \frac{|\{y\}^{\downarrow} \cap A|}{|A|} \cdot \frac{|(X - A) - \{y\}^{\downarrow}|}{|X - A|}. \tag{3.14}$$

In this case, the second factor, $\frac{|(X-A)-\{y\}^{\downarrow}|}{|X-A|}$, is interpreted as the degree to which the objects outside of A do not have y.

A difference between (3.13) and (3.14) lays in (3.13) considering objects that belong to both A and $\{y\}^{\downarrow}$ twice, while (3.14) considering each object only once, the first time in A, the second time in $X - A$. In our experiments, we used the first formula (3.13) because it yields a slightly better performance of typicality prediction.

The attribute weight proposed above can be used in the typicality formula:

Definition 9. *The degree $typ_w(x, A)$ of typicality of an object x in A is then defined by*

$$typ_w(x, A) = \sum_{y \in \{x\}^{\uparrow}} w(y, A), \tag{3.15}$$

where the attribute weight $w(y, A)$ is defined as the characteristicness of y with respect to A by (3.13).

3.3 Conclusions

We have described two basic schemes of typicality based on Rosch and Mervis' work: the original formula based on verbal description and a more general similarity-based scheme.

We have then shown that the Rosch and Mervis' formula can be extended to consider the absence of the attributes. Remarkably, examining the mutual relationship between these two approaches shows that the original and extended formulas are equal to the scaled similarity-based scheme with a specifically selected well-known similarity coefficient: Russel-Rao and simple matching coefficient.

Examination of the attribute weights used in the original Rosch and Mervis' formula led us to the proposition of a completely new attribute weight inspired by work related to characteristic attributes of objects, which takes the outside of the category into account.

Chapter 4

Experimental evaluation

This chapter contains a high-level overview of experiments and their results conducted by Belohlavek and Mikula (2022, 2024a,b,c,d), which are also available as appendices A–E.

4.1 Datasets

The following section provides an overview of the two datasets used in our experimental evaluation.

4.1.1 Zoo dataset

The Zoo dataset (Forsyth, 1990) is a commonly known dataset that describes 101 animals (exemplars) by their 17 attributes. All attributes, except the “legs”, are binary (yes/no) attributes. For our purposes, the “legs” attribute was binarized by binary scaling. Thanks to its small size and binary nature, the Zoo can be considered a well-suited dataset for basic experiments. However, it is worth mentioning that the dataset contains rows with duplicated attributes corresponding to two distinct objects. For example, “bass”, “catfish”, “chub”, “herring”, and “piranha” are the same animals according to the Zoo dataset. This limitation results from a trade-off between the size of the dataset and its descriptive power, and it is worth keeping it in mind during the interpretation of the results.

Human judgment of typicality

We obtained human typicality ratings for selected categories from the Zoo data to perform comparative experiments. These categories are: “bird”, “fish”, and “mammal”.

The study was conducted on 242 respondents, approximately 56% (136) of women and 44% (106) of men. Participants’ median, minimum, and maximum age were 23, 17, and 81. Respondents were split into four groups (students at Palacky University Olomouc, our coworkers, relatives, and others), which allowed us to verify reliabilities between these groups. Since correlation analysis revealed high correlations between groups, data from different groups were merged into a single set.

All respondents filled out an online questionnaire that included all three categories with corresponding exemplars: “bird” (20 exemplars), “fish” (13), and “mammal” (41). The typicality of every exemplar was rated on a scale from 1 (the least typical) to 5 (the most typical). Note that each participant’s order of objects within the category was randomized. The final typicality rating for each object from the category was obtained by calculating the mean value of all respondents’ ratings (Belohlavek and Mikula, 2021).

4.1.2 Dutch data

This section provides a basic overview of the Dutch data[1] and its parts relevant to the following sections. Because of its psychological significance, we encourage the study of the original publication for future details and insights (De Deyne et al., 2008; Ruts et al., 2004).

The data published in “Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts” (De Deyne et al., 2008) are unique in quality and size. To our knowledge, these data are the most comprehensive regarding common human categories and their numerous characteristics.

The Dutch data are based on studies involving hundreds of human respondents designed and gathered by psychologists at the University of Leuven. The data contains feature applicability matrices of objects (exemplars) and their attributes (features) from selected common linguistic categories alongside various psychologically relevant characteristics like typicality judgments and object similarity.

The dataset exemplars are members of 16 linguistic categories, including the natural kind and artifact categories (Ruts et al., 2004). There are 10 natural kind categories (249 exemplars total): “fruit” (30 exemplars); “vegetable” (30); “profession” (30); “sport” (30); the animal categories “amphibian” (5), “bird” (30), “fish” (23), “insect” (26), “mammal” (30), and “reptile” (22). The 6 artifact categories (166 exemplars total) are: “clothing” (29), “kitchen utensil” (33), “musical instrument” (27), “tool” (30), “vehicle” (30), and “weapon” (20).

Features for these exemplars were gathered in two ways. The category attributes were obtained by listing relevant attributes for the given category. On the other hand, the exemplar attributes were gathered by listing attributes for every exemplar from a given category. Unique sets of animal and artifact domain attributes were created by merging all attributes of categories from the corresponding domain.[2] Note that not all categories are included in some domain, e.g., the “sport” category is excluded from both.

The original data contains minor semantic and technical problems limiting machine processing. We thus modified the data to improve these aspects. These improvements include fixing misspelled/duplicated objects and attribute names in English, converting all names of objects and attributes into lowercase, transforming category names into singular, and fixing invalid CSV format of some files. The corrected version of Dutch data is publicly available on Github (Belohlavek and Mikula, 2022, 2023).

[1] We picked the name “Dutch data” for convenience, the full name of this dataset is “Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts”.

[2] Each artifact and animal domain has a category and exemplar-based attribute set similar to the original category features.

Exemplar by feature applicability matrices

The exemplar by feature applicability matrices are the core of Dutch data. These matrices contain information on how many respondents have agreed on whether an exemplar does or does not have a given attribute. Each category and domain has two matrices, one for category attributes and the other for exemplar attributes.

Each matrix was filled by 4 individual respondents, resulting in aggregated matrices that contain values from 0 to 4, indicating how many respondents found the agreement on exemplar having given attribute. These aggregated matrices can be transformed into binary matrices by thresholding. One could choose thresholds 2 or 3 (at least two/three respondents agreed) as relevant thresholds to avoid extreme thresholds like 1 (at least one respondent agreed) or 4 (all respondents agreed).

Typicality ratings

An essential part of the data for our research is the human typicality judgments. For each object from the 15 categories ("amphibian" category was excluded since these exemplars are included in "reptiles" category), typicality judgment on the scale of 1 (very atypical) to 20 (very typical) was obtained across 112 respondents. The mean typicality rating for each object was calculated as the mean value across all respondents who rated a given object.

Similarity judgements

Other important data are pairwise similarity judgments. These were obtained for each pair of objects from 15 categories ("amphibian" category was excluded since these exemplars are included in "reptiles" category) by 42 respondents in a previous study (Ruts et al., 2004) and extended later by another 92 respondents (De Deyne et al., 2008). Each pair of objects was rated on a scale of 1 (totally dissimilar) and 20 (totally similar).

Note that similarity was erroneously filled in the case of one respondent in the "fruit" category, resulting in asymmetrical similarity ratings. Since we cannot determine if this was caused by data postprocessing or the respondent itself, we omitted this data point in our modified version.

4.2 Experimental foundation

The following section describes the basic set of experiments conducted on a subset of categories ("bird", "fish", "mammal") to verify the feasibility of the typicality definition described in Chapter 2. The main goal was to measure the agreement between typicality calculated by our formulas and human judgment of typicality. These experiments were conducted firstly on the Zoo data (Belohlavek and Mikula, 2020) and later extended by experiments on the Dutch data (Belohlavek and Mikula, 2022). This section contains an overview of the results; see the Appendix A for a detailed discussion.

Degrees of typicality were calculated via formula (3.11) for each object per all available categories. Three selected similarity measures were tested: simple matching coefficient (SMC), Jaccard index (Jac), and Russel-Rao (RR) similarity. Recall that Russel Rao's similarity equals the similarity rating derived from Rosch and Mervis' formula for typicality. Thus, they provide the same resulting typicality order. The Kendall τ_b correlation coefficient was used to compare the calculated degrees of typicality and human typicality ratings. For each category, the Kendall τ_b was calculated for every pair of calculated typicality degree and human ratings.

Note first that according to a commonly accepted interpretation, the values of τ_b may be interpreted as follows: $\tau_b \geq 0.3$, $0.2 \leq \tau_b < 0.3$, $0.1 \leq \tau_b < 0.2$, and $0.0 \leq \tau_b < 0.1$ indicate strong, moderate, weak and very weak correlation, respectively (Belohlavek and Mikula, 2022).

In the case of Zoo data, most of the Kendall τ_b values were within the range of weak to moderate correlation with the exception of "fish" category, which shows no correlation for all tested similarities. We discovered that most of our respondents consider the "carp" as one of the most typical fish. Our hypothesis is the strong cultural influence of the Czech Republic. However, the attributes available in the Zoo dataset caused the "carp" to be an atypical fish.[3] To support this explanation, we repeated the experiments with the modified "fish" category (the "carp" was omitted). Modified "fish" category reached the expected correlation in the range of weak to moderate for all tested similarity measures.

Most of the Kendall τ_b values calculated on the Dutch data were higher than in the Zoo data, ranging between moderate and strong correlations. One notable exception was the "mammal" category, which exhibits a very weak correlation for SMC and Jac similarity coefficients, but a strong correlation in the case of the RR similarity coefficient. These results show the major influence of the difference between the size and quality of datasets.

These experiments provided promising results that motivated our extensive experimental research. At this point, one could ask if there are any similarity measures that provide a better degree of typicality prediction concerning human typicality judgments. This question will be experimentally examined in the next section.

4.3 Similarity measures and their influence

After promising results from the basic set of experiments, we examined 69 similarity measures alongside human similarity ratings and their influence on typicality degree calculation (Belohlavek and Mikula, 2024d). The full paper is available as Appendix B.

The experiment design was similar to the one from the previous section. We increased not only the number of tested similarity measures but also the number of categories. We included all available categories from both animal and artifact domains.

Since Dutch data includes human similarity ratings HJ for any pair (x, y) of objects from any given category (see Section 4.1.2), we used these values to build a similarity function sim_{HJ}, which was later used in the calculation of typicality degrees.[4] Surprisingly, there were better similarity functions to predict typicality in the given categories than the human similarity judgments. We found a set of similarity measures that provide

[3] Mainly because it was the only fish with the presence of attribute "domestic" and absence of attribute "predator".

[4] Feature matrices are not used since all data about the exemplars are available as human similarity judgments.

a slightly better correlation. Namely Co1, RR, int, Di2, and CT3.[5] Nevertheless, the typicality degree derived from HJ similarity still provides a strong Kendall τ_b correlation of value 0.42.

A closer examination found that typicality based on HJ performed better in the context of the animal domain. On the other hand, similarities Co1, RR, int, Di2, and CT3 performed better in artificial domain categories. One possible explanation is that the similarity of animal exemplars is more straightforward to judge than the similarity of objects from the artifact domain. The second possible explanation is that features provided in the animal domain are more descriptive than features of the artifact domain.

Other groups of similarities can be identified in the experimental results. Group which performed slightly worse than HJ which includes Fai, FM, CT4, Fos, Ku2, McC, Sor, SS1, cos, Jac, Maa, and Gle. One can also see similarities with a high average correlation except category attributes data, these are Den, Co2, Col, Di1, Twd, Fo1, and Gow. The number of relevant similarity measures that can be considered well-suited for predicting typicality ratings is surprising.

Overall, these experimental findings support the similarity-based scheme presented in Section 3.2 and provide a new way of comparing and benchmarking similarity measures.

4.4 Characteristic attributes

After examining the similarity-based scheme, we asked if there is an even better way to predict the typicality ratings. We focused on improving the attribute weights described in Section 3.2.5 and experimentally examined the plausibility of the newly proposed typ_w scheme. We compared the typ_w typicality ratings with the top 10 similarity-based typicality ratings examined in the previous section. These included typicality ratings based on similarity measures: Co1, RR, int, Di2, and CT3 alongside the HJ human similarity ratings.

One may explore multiple questions regarding the newly proposed formula. Our first question was if the typ_w is strictly better than the similarity-based scheme. We found that in the average case, the typ_w provides a better prediction of typicality across both domains. This is a remarkable achievement since we could not significantly overcome the similarity-based scheme based on the Russel-Rao similarity coefficient before.

Secondly, we examined the possible influence of a particular domain, which was previously observed as a significant factor in predicting power. We found that in the case of the animal domain, typ_w outperformed the previously best typicality prediction based on human similarity ratings typ_{HJ}. We found it especially important because we initially thought that the animal domain suffers from the consequences of human respondents not adequately describing the animal exemplars with attributes. Thus, the typicality calculated via these attributes will consistently outperform the typicality based on human similarity ratings. This is even more significant if we consider the gap between the typicality based on HJ and the rest of the similarity measures.

Last but not least, we investigated the influence of attribute type. Let us recall that we found support for the influence of category and exemplar-based attributes in our previous

[5] Full name of the similarity measure and their formal definition is available in the corresponding paper.

studies. Surprisingly, we found the impact of attribute type to be minimal. We explained it by the fact that domain data describe the category members in the broader context, where the difference between category and exemplar attributes is minimal.

In conclusion, we found significant experimental support for the newly proposed formula typ_w. Experimental results support the hypothesis that typicality prediction can be improved further.

4.5 Similarity and human judgment

Section 4.3 described how well the similarity measures suit typicality degree calculation. In this section, we will describe our approach to a direct comparison of similarity coefficients to the human similarity ratings available in the Dutch data (Belohlavek and Mikula, 2024b). The full paper is available as Appendix D.

The rationale of our experiments was as follows. Since the Dutch data includes human similarity rating pairs $\mathrm{HJ}(x, y)$ alongside feature matrices with their attributes (see Section 4.1.2), we calculated the similarity degree of x and y for each of the 69 similarity measures and compared these to HJ. Therefore, we got 69 $sim(x, y)$ values for every possible pair of objects x, y in each available category. These pairs of objects were ordered according to their $sim(x, y)$ value from the most to the least similar. These orderings were later compared with orders based HJ via Kendall τ_b order rank correlation. We found that, except for a few cases, similarity measures strongly correlate to the similarity based on human similarity ratings.

Our experiments revealed interesting results regarding attribute types. The HJ similarity is solely based on how human respondents rated the similarity of the two exemplars. On the other hand, the remaining similarity measures were tested on four types of feature matrix data (see Section 4.1.2). We found that similarity values calculated on the exemplar attributes provide stronger correlations than category attributes. Namely, the correlations were stronger on category-based data with exemplar attributes, followed by domain-based data with exemplar attributes, domain-based data with category attributes followed with a surprisingly small gap, and the worst performance was observed on category-based data with category attributes.

Another question we addressed in our experiments is the role of the shared absence. One significant question in the literature on similarity measures is whether the shared absence of attributes should contribute to the final similarity value of two objects. Our experiments supported the hypothesis that shared absence does not significantly contribute to the similarity of two objects since similarities without shared absence yielded a stronger correlation to the HJ similarity.

We also examined the perfect correlated similarities since multiple similarity measures provided identical orderings even if they provided different similarity values. Lastly, we compared possible similarity grouping with the groups from other comparative literature. Surprisingly, the resulting groupings were similar to those based on different authors' datasets.

4.6 Are human categories formal concepts?

One could ask how well a formal concept analysis models the real-life categories in real data (Belohlavek and Mikula, 2024a). We asked a question about how many of the categories form formal concepts in corresponding formal contexts. This section provides an overview of our findings. The full text is available as Appendix E.

Each of the 16 feature matrices for the animal and artifact domains obtained from the original feature matrices by thresholds "≥ 1," "≥ 2," "≥ 3," and "$= 4$," was counted as a formal context with corresponding concept lattice $\mathcal{B}(X, Y, I)$. These formal contexts were used to test if the given category is equal to the extent of the corresponding formal concept.

Each test determined whether the given category is a formal concept under a given feature matrix with a corresponding threshold. This led to 88 possible tests. The 63 of these 88 tests concluded that category is, in fact, a formal concept (roughly 71% of tests were positive). When we broke down the test results between the animal and artifact domain, we got 37 positive tests from 40 total tests corresponding to the animal domain and 26 positive tests from 48 total of the artifact domain. This notable difference between natural and artifact categories supports previous research from the psychology field (Belohlavek and Mikula, 2024a).

These results are impressive, mainly if we consider the high sensitivity of the notion of formal concept to the attributes of objects. A single attribute can influence whether the category represents a formal concept. When we considered the most meaningful thresholds "≥ 2," and "≥ 3," we got 34 positive tests from a total of 44 (roughly 77% of tests were positive).

We also conducted similar tests on the feature matrices from particular respondents to compare the difference between a person's and consensus knowledge. In this case, feature matrices naturally form binary data, and thresholding was not required. Almost all (39 of 40) tested categories from the animal domain form formal concepts across respondents. On the other hand, only 26 of 48 tests on the artifact domain confirmed the presence of a formal concept. These results again confirmed the notable difference between the animal and artifact domains.

4.7 Conclusions

We have described Zoo and Dutch data used in multiple experiments. The experimental foundation was laid off by examining the Rosch and Mervis' scheme alongside the similarity-based scheme on the Zoo and Dutch datasets. Since we found promising results, we did much larger experiments on the Dutch dataset. We have examined the performance of 69 similarity measures and the human similarity judgment provided in the Dutch data. Then, we have shown that the newly proposed typicality scheme with a new weight based on the characteristic attributes outperforms all previously examined approaches to typicality.

Dutch data were used not only to test the typicality formalization but also to directly compare human similarity judgments from Dutch data with the 69 similarity measures we carefully gathered from multiple studies. We found evidence that most similarity measures are more or less adequate compared to human judgments of similarity. Moreover, we found a similar grouping of similarity measures to those of other comparative studies.

And lastly, we examined the Dutch data categories in the light of formal concept analysis. Surprisingly, we found that the majority of categories presented in the Dutch data are, in fact, formal concepts. This result is interesting mainly in the context of classical theory of concepts which is considered obsolete in the psychology of concepts.

Chapter 5

Conclusions

This thesis summarized the typicality phenomenon in the context of the psychology of concept and presented the possible approaches to its computational account. We briefly described the history of theories of human concepts, which have shown the importance of the human conceptual system, and discussed possible pitfalls during its research. The wide range of theories suggests the idea that the "correct answer" is still far away in the future.

Behavioral observations and determinants of typicality were summarized, and we have described that typicality has the strongest and most reliable effects across multiple fields. Its influence can be seen in category judgment, category learning, category inference, exemplar generation frequency, linguistics, and many more. Rosch and Mervis proposed the family resemblance hypothesis, which changed the approach to understanding the structure of categories. Barsalou demonstrated that family resemblance is not the only determinant of typicality and that typicality can be highly dynamic depending on the context. More recently, Dieciuc and Folstein proposed the structural and functional typicality framework supported by neuroimaging and behavioral research, which provided a unified ground for examining stable and dynamic effects of the typicality phenomenon.

We described two basic formalizations of typicality based on Rosch and Mervis' work: the original formula based on the verbal description and a more general similarity-based scheme we proposed. We have shown that Rosch and Mervis' formula can be extended to consider the absence of the attributes. Remarkably, examining the mutual relationship between these two approaches showed that the original and extended formulas equal the scaled similarity scheme with a specifically selected well-known similarity measure: Russel-Rao and simple matching coefficient. Examination of the attribute weights used in the original Rosch and Mervis' formula led us to the proposition of a completely new attribute weight inspired by work related to characteristic attributes of objects, which takes the outside of the category into account.

These formalizations were tested using a set of experiments conducted on the two datasets. The first set of experiments was done on the Zoo data and a subset of Dutch data. Since Zoo data does not include human typicality ratings, we gathered typicality ratings from 242 respondents. The second set of experiments was conducted on a much larger dataset called Dutch data. We have examined the performance of 69 similarity measures and human similarity judgments provided as part of the Dutch data. The third set of experiments showed that the newly proposed typicality scheme with a new weight based on the characteristic attributes outperforms all previously examined approaches to the typicality.

Dutch data were used to test the formalization of typicality and to directly compare human similarity judgments from Dutch data to the 69 similarity measures we carefully gathered from multiple studies. We found evidence that most similarity measures are more or less adequate compared to human judgments of similarity. Moreover, other comparative studies found a similar grouping of similarity measures.

And lastly, we examined the Dutch data categories in the light of formal concept analysis. Surprisingly, we found that the majority of categories presented in the Dutch data are, in fact, formal concepts. This result is interesting mainly in the context of classical theory of concepts which is considered obsolete in the psychology of concepts.

Overall, we have found solid support for multiple schemes of typicality, which includes completely new approaches. This can motivate future research, giving even more insight into this fundamental phenomenon. As the modeling effort brings more detailed models of human cognition, it opens the possibilities to evaluate and examine these theories. Most experiments done with human respondents can be executed and evaluated with large language models, hopefully bringing more interesting insights into human cognition.

Shrnutí v českém jazyce

V této práci jsme shrnuli fenomén typičnosti v kontextu kognitivní psychologie a představili jsme možné přístupy k jeho výpočetnímu zpracování. Stručně jsme popsali historii teorií lidských konceptů a diskutovali možné nástrahy během jeho výzkumu. Široká škála teorií naznačuje, že správná odpověď je stále daleko v budoucnosti.

Shrnuli jsme behaviorální pozorování a determinanty typičnosti a ukázali jsme, že typičnost má jeden z nejsilnějších vlivů napříč mnoha oblastmi. Její vliv lze pozorovat v kategorizaci, učení, inferenci, frekvenci generování exemplářů, lingvistice a mnoha dalších oblastech. Rosch a Mervis navrhly hypotézu family resemblance, která změnila přístup k chápání struktury kategorií. Barsalou ukázal, že family resemblance není jediným determinantem typičnosti a že typičnost může být vysoce dynamická v závislosti na kontextu. V nedávné době Dieciuc a Folstein navrhli teorii strukturální a funkční typičnosti, která poskytuje jednotný základ pro zkoumání stabilních a dynamických účinků fenoménu typičnosti a je podpořena neurozobrazovacími a behaviorálními výzkumy.

Dále jsme popsali dvě základní formalizace typičnosti založené na práci Rosch a Mervis: původní formuli založenou na slovním popisu a obecnější schéma založené na měrách podobnosti. Ukázali jsme, že formuli od Rosch a Mervis lze rozšířit o zohlednění absence atributů. Zkoumání vzájemného vztahu mezi těmito přístupy ukázalo, že původní a rozšířené formule jsou při vhodném výběru míry podobnosti rovny škálovanému schématu založnému na podobnosti. Zkoumání vah atributů používaných v původní formuli od Rosch a Mervis nás vedlo k návrhu zcela nové váhy atributů inspirované prací související s charakteristickými atributy objektů, která zohledňuje vnější část kategorie.

Tyto formalizace byly testovány pomocí sady experimentů provedených na dvou datových sadách. První sada experimentů byla provedena na Zoo datech a malé podmnožině Dutch dat. Vzhledem k tomu, že Zoo data nezahrnují lidská hodnocení typičnosti, shromáždili jsme hodnocení typičnosti od 242 respondentů. Druhá sada experimentů byla provedena na celých Dutch datech. Zkoumali jsme výkon 69 měr podobnosti a lidských hodnocení podobnosti poskytnutých jako součást Dutch dat. Třetí sada experimentů ukázala, že nově navržené schéma typičnosti s novou váhou založenou na charakteristických atributech překonává všechny dříve námi zkoumané přístupy k typičnosti.

Dutch data byla použita nejen k testování formalizace typičnosti. Rovněž jsme provedli přímé srovnaní hodnocení podobnosti od respondentů z Dutch dat s 69 míry podobnosti, které jsme pečlivě shromáždili z několika studií. Zjistili jsme, že většina měr podobnosti je více či méně srovnatelná s lidským hodnocením podobnosti.

Nakonec jsme zkoumali kategorie Dutch dat ve světle formální konceptuální analýzy. Překvapivě jsme zjistili, že většina kategorií prezentovaných v Dutch datech jsou ve skutečnosti formální koncepty. Tento výsledek je zajímavý především v kontextu klasické teorie konceptů, která je považována za zastaralou.

Výsledky naši práce poskytují pevný základ pro několik schémat typičnosti, které zahrnují zcela nové přístupy a poznatky. Tyto výsledky mohou motivovat budoucí výzkum, který poskytne ještě více poznatků o fenoménu typičnosti. Práce na modelech lidské kognice otevírá nové možnosti zkoumání a evaluaci těchto teorií. Většina experimentů provedených s lidskými respondenty může být dnes zopakována a vyhodnocena s pomocí velkých jazykových modelů, což může přinést zajimavé poznání o lidském konceptuálním systému.

Bibliography

Ahn, W.-k. et al. (2000). “Causal Status as a Determinant of Feature Centrality”. In: *Cognitive Psychology* 41.4, pp. 361–416.

Atkinson, R. C. and R. M. Shiffrin (1968). “Human Memory: A Proposed System and Its Control Processes”. In: *Psychology of Learning and Motivation*. Ed. by K. W. Spence and J. T. Spence. Vol. 2. Academic Press, pp. 89–195.

Bailenson, J. N. et al. (2002). “A Bird’s Eye View: Biological Categorization and Reasoning within and across Cultures”. In: *Cognition* 84.1, pp. 1–53.

Barclay, J. R. et al. (1974). “Comprehension and Semantic Flexibility”. In: *Journal of Verbal Learning and Verbal Behavior* 13.4, pp. 471–481.

Barsalou, L. W. (1983). “Ad Hoc Categories”. In: *Memory & Cognition* 11.3, pp. 211–227.

Barsalou, L. W. (1985). “Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure in Categories”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11, pp. 629–654.

Barsalou, L. W. (1987). “The Instability of Graded Structure: Implications for the Nature of Concepts”. In: *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Emory Symposia in Cognition, 1. New York, NY, US: Cambridge University Press, pp. 101–140.

Barsalou, L. W. (2012). “The Human Conceptual System”. In: *The Cambridge Handbook of Psycholinguistics*. Ed. by M. Spivey, K. McRae, and M. Joanisse. 1st ed. Cambridge University Press, pp. 333–350.

Barsalou, L. W. (2015). “Situated Conceptualization: Theory and Applications”. In: *Perceptual and Emotional Embodiment*. Routledge.

Barsalou, L. W., L. Dutriaux, and C. Scheepers (2018). “Moving beyond the Distinction between Concrete and Abstract Concepts”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1752, p. 20170144.

Barsalou, L. W., W. Kyle Simmons, et al. (2003). “Grounding Conceptual Knowledge in Modality-Specific Systems”. In: *Trends in Cognitive Sciences* 7.2, pp. 84–91.

Barsalou, L. W. and D. R. Sewell (1984). “Constructing Representations of Categories from Different Points of View”. In: *Emory Cognition Project Technical Report #2*.

Belohlavek, R. (2008). *Introduction to Formal Concept Analysis*. Vol. 47.

Belohlavek, R. and T. Mikula (2020). “Typicality in Conceptual Structures within the Framework of Formal Concept Analysis”. In: CLA 2020.

Belohlavek, R. and T. Mikula (2021). *Zoo Typicality*. URL: `https://github.com/mikulatomas/zoo-typicality`.

Belohlavek, R. and T. Mikula (2022). “Typicality: A Formal Concept Analysis Account”. In: *International Journal of Approximate Reasoning* 142, pp. 349–369.

Belohlavek, R. and T. Mikula (2023). *Dutch Concepts*. URL: `https://github.com/mikulatomas/dutch-concepts`.

Belohlavek, R. and T. Mikula (2024a). "Are Human Categories Formal Concepts? A Case Study Using Dutch Data". In: *International Journal of General Systems*, pp. 1–50.

Belohlavek, R. and T. Mikula (2024b). *Comparing Similarity Measures for Binary Data with Human Judgment of Similarity*. Pre-published.

Belohlavek, R. and T. Mikula (2024c). *Schemes to Compute Typicality: Revisiting Rosch and Mervis's Approach*. Pre-published.

Belohlavek, R. and T. Mikula (2024d). "Similarity Metrics vs Human Judgment of Similarity for Binary Data: Which Is Best to Predict Typicality?" In: *Applied Soft Computing* 153, p. 111270.

Belohlavek, R. and M. Trnecka (2020a). "Basic Level of Concepts and Formal Concept Analysis 2: Examination of Existing Basic Level Metrics". In: *International Journal of General Systems* 49.7, pp. 707–723.

Belohlavek, R. and M. Trnecka (2020b). "Basic Level of Concepts in Formal Concept Analysis 1: Formalization and Utilization". In: *International Journal of General Systems* 49.7, pp. 689–706.

Blomberg, S. N. F. et al. (2019). "Injury from Electric Scooters in Copenhagen: A Retrospective Cohort Study". In: *BMJ Open* 9.12, e033988.

Blunden, A. (2012). *Concepts: A Critical Approach*. BRILL. 319 pp.

Borghesani, V. and M. Piazza (2017). "The Neuro-Cognitive Representations of Symbols: The Case of Concrete Words". In: *Neuropsychologia*. Special Issue: Concepts, Actions and Objects: Functional and Neural Perspectives 105, pp. 4–17.

Bozeat, S. et al. (2003). "A Duck with Four Legs: Investigating the Structure of Conceptual Knowledge Using Picture Drawing in Semantic Dementia". In: *Cognitive Neuropsychology* 20.1, pp. 27–47.

Brooks, L. R. (1978). "Nonanalytic Concept Formation and Memory for Instances". In: *Cognition and Categorization*. Ed. by E. Rosch and B. Lloyd. Lawrence Elbaum Associates, pp. 3–170.

Burnett, R. C. et al. (2005). "Ideal Is Typical". In: *Canadian Journal of Experimental Psychology* 59, pp. 3–10.

Carey, S. (1987). *Conceptual Change In Childhood*. Ed. by J. S. Rosenschein. Cambridge, Mass.: MIT Press. 240 pp.

Carpineto, C. and G. Romano (2004). *Concept Data Analysis: Theory and Applications*. John Wiley & Sons. 224 pp.

De Deyne, S. et al. (2008). "Exemplar by Feature Applicability Matrices and Other Dutch Normative Data for Semantic Concepts". In: *Behavior Research Methods* 40.4, pp. 1030–1048.

Dieciuc, M. A. and J. R. Folstein (2019). "Typicality: Stable Structures and Flexible Functions". In: *Psychonomic Bulletin & Review* 26.2, pp. 491–505.

Eysenck, M. W. and M. T. Keane (2020). *Cognitive Psychology: A Student's Handbook*. 8th ed. London: Psychology Press. 980 pp.

Folstein, J. R. and M. A. Dieciuc (2019). "The Cognitive Neuroscience of Stable and Flexible Semantic Typicality". In: *Frontiers in Psychology* 10.

Forsyth, R. (1990). *Zoo*. UCI Machine Learning Repository.

Frake, C. O. (2012). "The Ethnographic Study of Cognitive Systems". In: *The Ethnographic Study of Cognitive Systems*. De Gruyter Mouton, pp. 434–446.

Freeman, J. B. (2014). "Abrupt Category Shifts during Real-Time Person Perception". In: *Psychonomic Bulletin & Review* 21.1, pp. 85–92.

Ganter, B. and R. Wille (2012). *Formal Concept Analysis: Mathematical Foundations*. Springer Science & Business Media. 289 pp.

Garcez, A. d'Avila and L. C. Lamb (2020). "Neurosymbolic AI: The 3rd Wave". Prepublished.

Garrard, P. et al. (2001). "Prototypicality, Distinctiveness, and Intercorrelation: Analyses of the Semantic Attributes of Living and Nonliving Concepts". In: *Cognitive Neuropsychology* 18.2, pp. 125–174.

Garrod, S. and A. Sanford (1977). "Interpreting Anaphoric Relations: The Integration of Semantic Information While Reading". In: *Journal of Verbal Learning and Verbal Behavior* 16.1, pp. 77–90.

Goodman, S. N., D. Fanelli, and J. P. A. Ioannidis (2016). "What does research reproducibility mean?" In: *Science Translational Medicine* 8.341, 341ps12.

Griffiths, T. L. (2015). "Manifesto for a New (Computational) Cognitive Revolution". In: *Cognition*. The Changing Face of Cognition 135, pp. 21–23.

Hampton, J. A. (1979). "Polymorphous Concepts in Semantic Memory". In: *Journal of Verbal Learning and Verbal Behavior* 18.4, pp. 441–461.

Hampton, J. A. (1981). "An Investigation of the Nature of Abstract Concepts". In: *Memory & Cognition* 9.2, pp. 149–156.

Hampton, J. A. (1982). "A Demonstration of Intransitivity in Natural Categories". In: *Cognition* 12.2, pp. 151–164.

Hampton, J. A., D. Dubois, and W. Yeh (2006). "Effects of Classification Context on Categorization in Natural Categories". In: *Memory & Cognition* 34.7, pp. 1431–1443.

Hinton, G. E. (1986). "Learning Distributed Representations of Concepts". In: *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Vol. 1. Amherst, MA, p. 12.

Hinton, G. E. and J. A. Anderson (1989). *Parallel Models of Associative Memory*. USA: L. Erlbaum Associates Inc.

Hsu, N. S., M. L. Schlichting, and S. L. Thompson-Schill (2014). "Feature Diagnosticity Affects Representations of Novel and Familiar Objects". In: *Journal of Cognitive Neuroscience* 26.12, pp. 2735–2749.

Hull, C. L. (1920). "Quantitative Aspects of Evolution of Concepts: An Experimental Study". In: *Psychological Monographs* 28, pp. i–86.

Chin-Parker, S. and B. H. Ross (2004). "Diagnosticity and Prototypicality in Category Learning: A Comparison of Inference Learning and Classification Learning". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.1, pp. 216–226.

Inhelder, B. and P. Jean (1964). *The Early Growth of Logic in the Child: Classification and Seriation*. The Early Growth of Logic in the Child: Classification and Seriation. Harper and Row: New York.

Iordan, M. C. et al. (2016). "Typicality Sharpens Category Representations in Object-Selective Cortex". In: *NeuroImage* 134, pp. 170–179.

Jackson, R. L., T. T. Rogers, and M. A. Lambon Ralph (2021). "Reverse-Engineering the Cortical Architecture for Controlled Semantic Cognition". In: *Nature Human Behaviour* 5.6 (6), pp. 774–786.

Kelly, M. H., J. K. Bock, and F. C. Keil (1986). "Prototypicality in a Linguistic Context: Effects on Sentence Structure". In: *Journal of Memory and Language* 25, pp. 59–74.

Kim, S. and G. L. Murphy (2011). "Ideals and Category Typicality". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37, pp. 1092–1112.

Lake, B. M., R. Salakhutdinov, and J. B. Tenenbaum (2015). "Human-Level Concept Learning through Probabilistic Program Induction". In: *Science* 350.6266, pp. 1332–1338.

Lake, B. M., T. D. Ullman, et al. (2017/ed). "Building Machines That Learn and Think like People". In: *Behavioral and Brain Sciences* 40, e253.

Lynch, E. B., J. D. Coley, and D. L. Medin (2000). "Tall Is Typical: Central Tendency, Ideal Dimensions, and Graded Category Structure among Tree Experts and Novices". In: *Memory & Cognition* 28.1, pp. 41–50.

Machery, E. (2009). *Doing without Concepts*. Oxford, New York: Oxford University Press. 296 pp.

Margolis, E. and S. Laurence, eds. (1999). *Concepts: Core Readings*. Cambridge, Mass: MIT Press. 652 pp.

Mayberry, E. J., K. Sage, and M. A. L. Ralph (2011). "At the Edge of Semantic Space: The Breakdown of Coherent Concepts in Semantic Dementia Is Constrained by Typicality and Severity but Not Modality". In: *Journal of Cognitive Neuroscience* 23.9, pp. 2240–2251.

McCloskey, M. E. and S. Glucksberg (1978). "Natural Categories: Well Defined or Fuzzy Sets?" In: *Memory & Cognition* 6.4, pp. 462–472.

Medin, D. L. and S. Atran (2004). "The Native Mind: Biological Categorization and Reasoning in Development and across Cultures". In: *Psychological Review* 111.4, pp. 960–983.

Medin, D. L. and M. M. Schaffer (1978). "Context Theory of Classification Learning". In: *Psychological Review* 85, pp. 207–238.

Mervis, C. B., J. Catlin, and E. Rosch (1976). "Relationships among Goodness-of-Example, Category Norms, and Word Frequency". In: *Bulletin of the Psychonomic Society* 7.3, pp. 283–284.

Mervis, C. B. and J. R. Pani (1980). "Acquisition of Basic Object Categories". In: *Cognitive Psychology* 12.4, pp. 496–522.

Murphy, G. (2004). *The Big Book of Concepts*. MIT Press. 574 pp.

Murphy, G. L. and D. L. Medin (1985). "The Role of Theories in Conceptual Coherence". In: *Psychological Review* 92, pp. 289–316.

Ojalehto, B. L. and D. L. Medin (2015). "Perspectives on Culture and Concepts". In: *Annual review of psychology* 66, pp. 249–275.

Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science". In: *Science* 349.6251, aac4716.

Osherson, D. N. et al. (1990). "Category-Based Induction". In: *Psychological Review* 97, pp. 185–200.

Patterson, K., P. J. Nestor, and T. T. Rogers (2007). “Where Do You Know What You Know? The Representation of Semantic Knowledge in the Human Brain”. In: *Nature Reviews Neuroscience* 8.12 (12), pp. 976–987.

Pobric, G., E. Jefferies, and M. A. Lambon Ralph (2010). “Category-Specific versus Category-General Semantic Impairment Induced by Transcranial Magnetic Stimulation”. In: *Current Biology* 20.10, pp. 964–968.

Posner, M. I. and S. W. Keele (1968). “On the Genesis of Abstract Ideas”. In: *Journal of Experimental Psychology* 77, pp. 353–363.

Posner, M. I. and S. W. Keele (1970). “Retention of Abstract Ideas”. In: *Journal of Experimental Psychology* 83, pp. 304–308.

Ralph, M. A. L. et al. (2017). “The Neural and Computational Bases of Semantic Cognition”. In: *Nature Reviews Neuroscience* 18.1 (1), pp. 42–55.

Ramesh, A. et al. (2022). “Hierarchical Text-Conditional Image Generation with CLIP Latents”. Pre-published.

Rips, L. J. (1975). “Inductive Judgments about Natural Categories”. In: *Journal of Verbal Learning and Verbal Behavior* 14.6, pp. 665–681.

Rips, L. J., E. J. Shoben, and E. E. Smith (1973). “Semantic Distance and the Verification of Semantic Relations”. In: *Journal of Verbal Learning and Verbal Behavior* 12.1, pp. 1–20.

Rogers, T. T. and J. L. McClelland (2004). *Semantic Cognition: A Parallel Distributed Processing Approach.* MIT Press. 446 pp.

Rosch, E. (1975). “Cognitive Representations of Semantic Categories”. In: *Journal of Experimental Psychology: General* 104, pp. 192–233.

Rosch, E. and C. B. Mervis (1975). “Family Resemblances: Studies in the Internal Structure of Categories”. In: *Cognitive Psychology* 7.4, pp. 573–605.

Rosch, E., C. B. Mervis, et al. (1976). “Basic Objects in Natural Categories”. In: *Cognitive Psychology* 8.3, pp. 382–439.

Rosch, E., C. Simpson, and R. S. Miller (1976). “Structural Bases of Typicality Effects”. In: *Journal of Experimental Psychology: Human Perception and Performance* 2, pp. 491–502.

Roth, E. M. and E. J. Shoben (1983). “The Effect of Context on the Structure of Categories”. In: *Cognitive Psychology* 15.3, pp. 346–378.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). “Learning Representations by Back-Propagating Errors”. In: *Nature* 323.6088 (6088), pp. 533–536.

Rumelhart, D. E. and J. L. McClelland, eds. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models.* Cambridge, MA, USA: MIT Press. 530 pp.

Rumelhart, D. E., P. Smolensky, et al. (1986). “Schemata and Sequential Thought Processes in PDP Models”. In: *Parallel Distributed Processing: Explorations in the Microstructure, Vol. 2: Psychological and Biological Models.* Cambridge, MA, USA: MIT Press, pp. 7–57.

Ruts, W. et al. (2004). “Dutch Norm Data for 13 Semantic Categories and 338 Exemplars”. In: *Behavior Research Methods, Instruments, & Computers* 36.3, pp. 506–515.

Sloman, S. A. and W.-K. Ahn (1999). “Feature Centrality: Naming versus Imagining”. In: *Memory & Cognition* 27.3, pp. 526–537.

Sloman, S. A., B. C. Love, and W.-k. Ahn (1998). "Feature Centrality and Conceptual Coherence". In: *Cognitive Science* 22.2, pp. 189–228.

Smith, E. E. and D. L. Medin (2013). *Categories and Concepts*. Harvard University Press.

Solomon, K. O., D. L. Medin, and E. Lynch (1999). "Concepts Do More than Categorize". In: *Trends in Cognitive Sciences* 3.3, pp. 99–105.

Stanley, D. J. and J. R. Spence (2014). "Expectations for Replications: Are Yours Realistic?" In: *Perspectives on Psychological Science* 9.3, pp. 305–318.

Tversky, A. (1977). "Features of Similarity". In: *Psychological Review* 84, pp. 327–352.

Ward, T. B. et al. (2000). "Attribute Centrality and Imaginative Thought". In: *Memory & Cognition* 28.8, pp. 1387–1397.

Wille, R. (1982). "Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts". In: *Ordered Sets*. Ed. by I. Rival. NATO Advanced Study Institutes Series. Dordrecht: Springer Netherlands, pp. 445–470.

Wittgenstein, L. (1968). *Philosophical Investigations*. Oxford, England: Basil Blackwell.

Woollams, A. (2012). "Apples Are Not the Only Fruit: The Effects of Concept Typicality on Semantic Representation in the Anterior Temporal Lobe". In: *Frontiers in Human Neuroscience* 6.

Yeh, W. and L. W. Barsalou (2006). "The Situated Nature of Concepts". In: *The American Journal of Psychology* 119.3, pp. 349–384.

Appendix A

Typicality: A formal concept analysis account

Paper (Belohlavek and Mikula, 2022) describes the basics of the typicality phenomenon, formalization inside the formal concept analysis, and fundamental experimental examination. The results of these experiments are briefly described in Section 4.2. The paper resulted from joint research with my supervisor Radim Bělohlávek and was published in the International Journal of Approximate Reasoning.

International Journal of Approximate Reasoning 142 (2022) 349–369

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Typicality: A formal concept analysis account

Radim Belohlavek, Tomas Mikula

Department of Computer Science, Palacky University, Olomouc, 17. listopadu 12, CZ-771 46 Olomouc, Czech Republic

ARTICLE INFO

Article history:
Received 28 February 2021
Received in revised form 25 October 2021
Accepted 1 December 2021
Available online 3 December 2021

Keywords:
Typicality
Categorization
Concept
Formal concept analysis
Psychology of concepts

ABSTRACT

We examine typicality—a significant phenomenon accompanying human concepts—within the framework of formal concept analysis. Our aim is to formalize the notion of typicality within this framework and thus provide an operational definition. We review relevant aspects and main psychological explanations of typicality, and propose a formalization based on a view of typicality propounded in the seminal work by Eleanor Rosch et al. We also provide experimental evaluation of our approach and discuss ramifications of our findings and topics to be explored in the future.

1. Aims of our study

Concepts are at the center of human reasoning and are hence the subject of numerous explorations. Among them, psychological studies of concepts have a distinguished role. The psychology of concepts provides a number of interesting theories and experimental studies of various phenomena involving concepts. These are of interest not only for the domain of psychology itself but, naturally, also for other domains concerned with concepts, including numerous formal approaches to reasoning and information processing using concepts.

One of the most significant phenomena accompanying human concepts, which is broadly familiar from everyday life, is typicality: A sparrow is a typical bird, an ostrich is not; a trout is a typical fish, an eel or a flounder is not.[1] Typicality may be regarded as a manifestation of a graded structure of concepts, and plays a remarkable role in several cognitive tasks, such as categorization and classification, which are crucially important in processing information by humans.

Our aim in this paper is to formalize a view of typicality propounded in the seminal works on typicality and the graded structure of concepts by Eleanor Rosch et al. For our purpose, we utilize the framework of formal concept analysis (FCA).[2] This framework is naturally suited for our purpose because its fundamental notions, such as that of object, attribute, sharing of attributes, as well as other notions, appear as basic in most of the psychological studies of typicality. In a sense, selection of FCA represents a choice of a straightforward simple framework that provides formal counterparts to the primitive notions used informally by psychologists.

Two particular motivations for our study are as follows. First, formalization of typicality allows us to approach and explore typicality in precise terms amenable to formal analysis. This is important particularly in view of the fact that in the psychological literature, theories of typicality are described rather informally, very often just verbally. Similarly informal are reasoning and the conclusions regarding typicality presented in the literature. Formalization of typicality, on the other hand,

E-mail addresses: radim.belohlavek@acm.org (R. Belohlavek), mail@tomasmikula.cz (T. Mikula).

[1] In our cultural context; the role of cultural context is mentioned below.

[2] We thus continue our previous effort to examine the basic level of concepts [5–7] within FCA.

https://doi.org/10.1016/j.ijar.2021.12.001
0888-613X/

renders an operational definition which lets one realize the various subtleties and possible shortcomings of an informal, verbally described definition of typicality. In addition, it enables one to consider possible relationships to alternative definitions and related notions, and thus may generally help examine typicality in a more rigorous manner.[3]

Secondly, we believe that formalization of typicality is significant for FCA itself. In general, we consider extensions of data analytical and information processing methods, such as FCA, by notions coming from the psychology of concepts a meaningful task which may significantly enhance these methods. While typicality has as yet not been exploited in FCA, it seems a natural mean to extend the structure of formal concepts.

Our paper is meant to make first steps in studying typicality in the framework of FCA. In section 2, we provide an overview of typicality from the viewpoint of the psychology of concepts and present selected issues pertaining to typicality. Our formalization of typicality within the FCA framework is outlined in section 3. Examples and experiments involving typicality are the subject of section 4. Conclusions and a prospect of further topics to explore is outlined in section 5.

2. Psychological accounts of typicality

2.1. Typicality as manifestation of a graded structure of concepts

Until the 1970s, the prevalent paradigm in psychological studies of concepts was represented by the so-called classical view.[4] According to this view, a concept is determined by a set of yes/no (bivalent, binary) conditions (attributes, features) which are necessary and jointly sufficient, i.e. definitory in the following sense: An object is covered by (or, is a member of) the concept (or category in terms commonly used in the psychology of concepts) if and only if the object satisfies each of these conditions. This view has a long tradition in philosophy and logic and also underlies the notion of a concept in the basic setting of formal concept analysis.[5]

In the mid-1970s, various explorations—most importantly those led by Eleanor Rosch—in the internal structure of concepts revealed fundamental limitations of the classical view. Put briefly, it became apparent that concepts have a graded structure: Various phenomena had experimentally been found to be a matter of degree rather than bivalent (yes/no). In addition, important phenomena had been observed that were not accounted for by the classical view. Typicality, which is discussed in the first findings by Rosch et al. [22–24], represents such a phenomenon. Ever since these first findings, the phenomenon continues to be a subject of vivid psychological research; see e.g. [10,27].

The classical view does not account for typicality, at least not directly, which represents a shortcoming of this view. Namely, according to the classical view, all members of a category have an equal status with respect to the category. On the other hand, people naturally regard some objects more typical of a given category than other objects. Further research has shown that people are even capable of assigning degrees of typicality (called also typicality ratings) to objects for a given category in a consistent manner.

Note in this connection that another phenomenon, which had been examined in the early 1970s, that involves degrees and is not addressed by the classical view is the graded nature of a membership in category itself. That is, an object may not just be a member or a non-member of a given category, but rather a member to a certain degree in the sense of fuzzy sets.[6] While the classical view is constrained to two possible degrees of membership, namely 0 (non-member) and 1 (member), the more general view, which is experimentally confirmed as significantly more appropriate, allows for degrees of membership, such as 0.8 representing high but not full membership or 0.5 representing a borderline case.

Basically, there are two possible views to start from in considering typicality. The literature on the psychology of concepts does not, unfortunately, make it clear to which of these views a particular study of typicality subscribes; see. e.g. [20]. In the first view, membership in a category is bivalent (i.e. classical, yes/no) and typicality represents an additional structure of a category. In the second view, membership is graded and possibly even equivalent to (or otherwise strongly correlated with) typicality. In our formalization below we assume the former view, i.e. that categories (concepts) are classical and that typicality represents an additional structure. Such view is adopted, e.g., in the design of experiments in the seminal paper [23].

Note also an important feature of typicality, namely its high cognitive significance; see e.g. [1,20,23]. For one, people tend to agree on typicality ratings. Moreover, typicality is reported to predict performance in a variety of cognitive tasks including learning of categories (typical objects are learned more quickly), deciding membership in categories (decisions on typical objects are more quick), and production of category exemplars (typical exemplars are generated first). Typical items are also useful in making inferences about categories and serve as so-called cognitive reference points.

3 This aspect was a significant part of our work on basic level [5–7].

4 A detailed exposition of developments in the psychological theories of concepts is provided in the monograph [20]; see also [18].

5 Note, however, that in the fuzzy logic extension of the basic setting of formal concept analysis, attributes are considered fuzzy (graded) rather than bivalent; see e.g. [2,3].

6 Note that Rosch's studies of graded nature of categories were conducted independently and in about the same time as Zadeh's studies of fuzzy sets [28]. Both Rosch and Zadeh were with UC Berkeley.

2.2. Explanations of typicality

In their seminal paper [23], Rosch and Mervis put forward a hypothesis of what makes an object typical in a category. This hypothesis was confirmed by experiments by the authors [23] and had later been examined by numerous other studies; see e.g. the monograph [20], in which typicality occupies a significant part. Rosch and Mervis [23, p. 575] describe their hypothesis as follows:

> ...members of a category come to be viewed as prototypical of the category as a whole in proportion to the extent to which they bear a family resemblance to (have attributes that overlap those of) other members of the category. Conversely, items viewed as most prototypical of one category will be those with least family resemblance to or membership in other categories.

The first part referring to resemblance (similarity) to objects of the given concept (category) is intuitively compelling and relatively straightforward to formalize. It is this part that we use in our approach. The second part referring to resemblance to objects in other concepts is not so straightforward, brings non-trivial problems, which are also reflected in the experiments in [23], and we hence do not consider it in what follows.[7]

In addition, several other possible explanations of typicality of an item have been suggested and tested in later studies, including similarity to central tendency (central tendency being e.g. the average of a numerical characteristic of an item), closeness to ideals in goal-oriented categories (ideals represent characteristics that items should possess if they are to serve the goal associated with a category), frequency of instantiation (i.e. frequency of encounter with the item as a member of a given category), and familiarity (i.e. frequency of encounter across all contexts); see e.g. [1,19,20] and also [14]. A more recent research emphasizes also the role of context (situation) in which typicality is assessed [27]. The resulting instability of typicality resulting from dependence on context even led the authors in [10] to distinguish between the so-called structural typicality (representing stability) and functional typicality (representing context-dependence and thus instability).

In spite of several alternative hypotheses, the family resemblance hypothesis of Rosch and Mervis [23] mentioned above appears to remain the most simple and most commonly accepted explanation of typicality. It is due to this fact that Rosch and Mervis' explanation forms the basis of our approach to typicality.

3. Formalization of typicality within formal concept analysis

3.1. Preliminaries from formal concept analysis (FCA)

FCA [9,15] starts with its basic notion of a formal context, which is a triplet $\langle X, Y, I\rangle$ consisting of non-empty sets X and Y, and a binary relation (incidence relation) I between X and Y (that is, $I \subseteq X \times Y$, i.e. I consists of selected pairs $\langle x, y\rangle$). The sets X and Y are interpreted as the set of objects and the set of (yes/no) attributes, and the fact $\langle x, y\rangle \in I$ means that the object x has the attribute y. An example of a formal context is depicted in Table 5 in section 4: Objects $x \in X$ and attributes $y \in Y$ are represented by table rows and columns, and the incidence relation I by crosses and blanks; for $x =$ scorpion and $y =$ predator we have $\langle x, y\rangle \in I$ (scorpion is predator), for $x =$ frog and $y =$ hair we have $\langle x, y\rangle \notin I$ (frog does not have hair), etc.

A pair $\langle A, B\rangle$ consisting of a set $A \subseteq X$ of objects and a set $B \subseteq Y$ of attributes is called a formal concept in $\langle X, Y, I\rangle$ if and only if $A^{\uparrow} = B$ and $B^{\downarrow} = A$ where

$$A^{\uparrow} = \{y \in Y \mid \text{for each } x \in A : \langle x, y\rangle \in I\},$$
$$B^{\downarrow} = \{x \in X \mid \text{for each } y \in B : \langle x, y\rangle \in I\}.$$

Notice that $A^{\uparrow}$ and $B^{\downarrow}$ are the set of all attributes common to all objects in A and the set of all objects having all the attributes in B, respectively. Geometrically, formal concepts in $\langle X, Y, I\rangle$ are maximal rectangular areas (up to a permutation of rows and columns) in the table representing $\langle X, Y, I\rangle$ that are full of crosses. The notion of a formal concept corresponds to the traditional notion of concept as consisting of its extent (objects covered by the concept) and its intent (attributes characterizing the concept); the extent of a formal concept $\langle A, B\rangle$ is A, the intent is B.

In a given formal context $\langle X, Y, I\rangle$, there is, as a rule, a number of formal concepts. The set of all formal concepts in a given formal context $\langle X, Y, I\rangle$ is denoted by $\mathcal{B}(X, Y, I)$, i.e.

$$\mathcal{B}(X, Y, I) = \{\langle A, B\rangle \mid A^{\uparrow} = B \text{ and } B^{\downarrow} = A\},$$

and is called the concept lattice of $\langle X, Y, I\rangle$. Namely, when equipped with a natural subconcept-superconcept hierarchy $\leq$, defined by

[7] The problem is with the meaning of "other categories". We leave this problem for future research. Note, however, that the properties mentioned in the first part (i.e. similarity to objects of the given category, which we use) and the second part (small similarity to objects in other categories) were tested separately in [23], and that each of these two parts was found significantly correlated with typicality ratings.

$$\langle A, B\rangle \le \langle C, D\rangle \text{ if and only if } A \subseteq C \text{ (equivalently, if and only if } B \supseteq D),$$

the set $\mathcal{B}(X, Y, I)$ indeed becomes a complete lattice, whose structure is described by the basic theorem of concept lattices [15].

3.2. Our approach to typicality

As noted above, psychological explorations of typicality and other facets of the graded structure of concepts are considered a strong argument against the classical view of concepts. Since FCA is rooted in the classical view of concepts, one might conclude that using FCA is not appropriate for modeling typicality. In our view, this is not the case. We contend that typicality naturally occurs even in concepts with a yes/no membership resulting from yes/no attributes, as in the basic setting of FCA. Moreover, the seminal psychological experiments on typicality mentioned above, as well as several other studies of typicality in the psychological literature are based on the idea of objects described by yes/no attributes.

Let $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$ be a formal concept in a given formal context $\langle X, Y, I\rangle$. In accordance with Rosch and Mervis' view of typicality (section 2.2), we intend to regard an object x as typical for the given concept $\langle A, B\rangle$ to the extent to which it is similar to the objects in A, i.e. to the objects of this concept. A straightforward way is to assume a function

$$sim : X \times X \to [0, 1] \tag{1}$$

assigning to every two objects $x_1, x_2 \in X$ a number $sim(x_1, x_2) \in [0, 1]$ that may be interpreted as a degree to which x_1 and x_2 are similar (we come back to these functions below). Similarity of x to the objects x_1 in A, which underlies Rosch and Mervis' view of typicality, may naturally be interpreted as the average similarity of x to all the objects $x_1 \in A$. This leads to the following definition[8]:

Definition 1. Given a similarity (1), a *degree of typicality* of object $x \in A$ in a formal concept $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$ with $A \neq \emptyset$ is defined by

$$typ(x, \langle A, B\rangle) = \frac{\sum_{x_1 \in A} sim(x, x_1)}{|A|}. \tag{2}$$

Remark 1. (a) Admittedly, our approach is restrictive. One might, for instance, consider formula (2) for x not necessarily in A, or consider the notion of typicality of a subconcept, rather than an object, in a given concept. We proceed with our definition for simplicity.

(b) Typicality degrees provide additional information about a concept $\langle A, B\rangle$. Namely, they reveal a certain graded structure of the concept $\langle A, B\rangle$. Such a structure has a cognitive significance and may be further utilized. Notice that since $typ(x, \langle A, B\rangle) \in [0, 1]$ due to $sim(X, X) \subseteq [0, 1]$, the mapping $t : A \to [0, 1]$ defined by $t(x) = typ(x, \langle A, B\rangle)$ may be regarded as a fuzzy set [28] of objects typical of $\langle A, B\rangle$.

(c) The idea of an element being similar to other elements in a given set has been explored in the literature on clustering and machine learning in general; see e.g. [17,29] on typicality in clustering, and the literature on medoids in clustering, e.g. [21], and silhouettes in clustering, e.g. [25].

Let us now consider the choice of the similarity function (1). It seems natural to derive the degree $sim(x_1, x_2)$ to which the objects x_1 and x_2 are similar from the descriptions of these objects in terms of attributes, i.e. from the sets $\{x_1\}^\uparrow$ and $\{x_2\}^\uparrow$ (note that $\{x\}^\uparrow$ is the set of attributes possessed by x). We hence assume that

$$sim(x_1, x_2) = sim_Y(\{x_1\}^\uparrow, \{x_2\}^\uparrow), \tag{3}$$

where

$$sim_Y : 2^Y \times 2^Y \to [0, 1]$$

is a function assigning to arbitrary subsets B_1 and B_2 of the set Y of given attributes a degree $sim_Y(B_1, B_2) \in [0, 1]$ that may be interpreted as a degree of similarity of B_1 and B_2. Such functions have been studied in various areas, most notably in the field of clustering; see e.g. [13].

Two particular functions serving this purpose, which we use in our experiments, are the well-known Jaccard index [16], sim_{J}, and the simple matching coefficient, sim_{SMC}, defined by

[8] Average similarity is mentioned in some psychological studies; see e.g. [1, p. 630]. Note that we also explored minimum instead of average, as it represents the best lower similarity-threshold. Average, nevertheless, yielded more intuitive results. We use $[0, 1]$ for the range (i.e. similarity is scaled), but $\mathbb{R}^+$ is also a natural option (non-scaled).

$$sim_{\mathrm{J}}(B_1, B_2) = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|} \quad \text{and} \tag{4}$$

$$sim_{\mathrm{SMC}}(B_1, B_2) = \frac{|B_1 \cap B_2| + |Y - (B_1 \cup B_2)|}{|Y|}, \tag{5}$$

respectively. That is, $sim_{\mathrm{J}}(B_1, B_2)$ is the number of attributes that belong to both B_1 and B_2 divided by the number of all attributes that belong to B_1 or B_2; $sim_{\mathrm{SMC}}(B_1, B_2)$ is the number of attributes on which B_1 and B_2 agree (either $y \in B_1$ and $y \in B_2$, or $y \notin B_1$ and $y \notin B_2$) divided by the number of all attributes. Hence, while sim_{SMC} treats both presence and non-presence of attributes symmetrically, sim_{J} disregards non-presence. This is the main conceptual difference between sim_{J} and sim_{SMC}.

The choice of the similarity sim_Y is in a sense crucial and, obviously, several other options different from sim_{J} and sim_{SMC} are possible. In this study, we nevertheless refrain from exploiting the variety of possible further similarity functions. Note, however, that in the next section, we naturally come to a third similarity, which we consider in this paper.

3.3. Relationship to Rosch and Mervis's formula for typicality

Formula (2) for typicality derives in a straightforward (and–as we contend–the most direct) way from the verbal description of Rosch and Mervis's hypothesis quoted in section 2.2. Interestingly, in their experiments to test the hypothesis, Rosch and Mervis [23] use a different formula for typicality of an object. Strangely, this formula does not bear a direct connection to similarity of objects, which is crucial in the hypothesis. The formula is described in [23] as follows. Given a concept, one assigns to every attribute its weight, namely the number of all objects of the concept that possess the attribute. A typicality of a given object in the concept is then the sum of the weights of all the attributes possessed by the object.

This definition translates to the FCA framework as follows. For a given $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$ and $y \in Y$, put

$$w(y, \langle A, B\rangle) = |\{x \in A \mid x \in \{y\}^{\downarrow}\}| \quad \text{(weight of attribute } y\text{)}.$$

Now, according to Rosch and Mervis, the typicality of the object $x \in A$ with respect to the concept $\langle A, B\rangle$ is defined by

$$typ_{\mathrm{RM}}(x, \langle A, B\rangle) = \textstyle\sum_{y \in \{x\}^{\uparrow}} w(y, \langle A, B\rangle).$$

The following theorem shows that in fact, Rosch and Mervis's formula for typicality, which is on the first sight of a different sort compared to our (2), may actually be regarded as resulting from a particular case of our scheme (2) by a simple scaling.

Theorem 1. *For the function* $sim_{\mathrm{rm}}(x_1, x_2) = \frac{|\{x_1\}^{\uparrow} \cap \{x_2\}^{\uparrow}|}{|Y|}$ *we have*

$$typ_{\mathrm{RM}}(x, \langle A, B\rangle) = |A| \cdot |Y| \cdot typ_{\mathrm{rm}}(x, \langle A, B\rangle)$$

where $typ_{\mathrm{rm}}(x, \langle A, B\rangle)$ *is determined by* sim_{rm} *according to (2).*

Proof. Since

$$|A| \cdot |Y| \cdot typ_{\mathrm{rm}}(x, \langle A, B\rangle) = |A| \cdot |Y| \cdot \frac{\sum_{x_1 \in A} sim_{\mathrm{rm}}(x, x_1)}{|A|}$$

$$= |A| \cdot |Y| \cdot \frac{\sum_{x_1 \in A} \frac{|\{x\}^{\uparrow} \cap \{x_1\}^{\uparrow}|}{|Y|}}{|A|} = \sum_{x_1 \in A} |\{x\}^{\uparrow} \cap \{x_1\}^{\uparrow}|,$$

we clearly need to verify

$$typ_{\mathrm{RM}}(x, \langle A, B\rangle) = \sum_{x_1 \in A} |\{x\}^{\uparrow} \cap \{x_1\}^{\uparrow}|.$$

Denoting $||\varphi||$ the truth value of φ (e.g. $||y \in \{x_1\}^{\uparrow}|| = 1$ iff $y \in \{x_1\}^{\uparrow}$), we obtain

$$\sum_{x_1 \in A} |\{x\}^{\uparrow} \cap \{x_1\}^{\uparrow}| = \sum_{x_1 \in A} \sum_{y \in \{x\}^{\uparrow}} ||y \in \{x_1\}^{\uparrow}||$$

$$= \sum_{x_1 \in A} \sum_{y \in \{x\}^{\uparrow}} ||x_1 \in \{y\}^{\downarrow}|| = \sum_{y \in \{x\}^{\uparrow}} \sum_{x_1 \in A} ||x_1 \in \{y\}^{\downarrow}||$$

$$= \sum_{y \in \{x\}^{\uparrow}} |\{x_1 \in A \mid x_1 \in \{y\}^{\downarrow}\}| = \sum_{y \in \{x\}^{\uparrow}} w(y, \langle A, B\rangle) = typ_{\mathrm{RM}}(x, \langle A, B\rangle),$$

completing the proof. □

Remark 2. (a) For a given formal concept $\langle A, B\rangle$ there thus exists a constant c (namely, $c = |A| \cdot |Y|$) such that Rosch-Mervis typicality typ_{RM} is obtained as a c-multiple of a particular typicality (namely typ_{rm}) obtained from our scheme.

(b) As a consequence, the list of objects sorted by typ_{RM} coincides with the list sorted by typ_{rm}.

4. Experiments

We performed experiments with data reported in [11] (section 4.1) and the well-known Zoo data [12] (section 4.2). Our main goal was to observe whether our formulas for typicality agree with human judgment, that is observe to what extent the rankings of objects by typicality degrees (or typicality ratings in terms often used in the psychological literature) computed by our formulas agree with the rankings resulting from human judgment, i.e. human assessment of typicality, for the above data. While human judgment data are available for the data in [11], we had to obtain the human judgment data for the Zoo data by our own questionnaire. Secondly, we attempted to analyze relationships between the formulas for typicality, which we provided, by observing agreements of the typicality degrees computed by the formulas. For the purpose of observing such agreements, both between our formulas and human judgment and between pairs of our formulas, we utilized various rank order correlation coefficients.

4.1. Experiments with Dutch data

Dutch data and the parts used in our experiments

The data used in this section is presented in [11], a study which provides perhaps the most comprehensive data regarding common human categories and their numerous characteristics, including typicality ratings. We first provide a brief description of the data and describe which part we used; the reader is referred to [11] for details.

The data comprises information on both the so-called natural kind and artifact categories, as these two types of categories are believed to have distinct properties (such as mental representation) by the psychologists. The data includes 16 human categories, each of which is represented by a number of selected exemplars (i.e. objects in the sense of FCA). A set of exemplars for a given category is to be considered an extent of the category (we come later to whether it actually is an extent in the sense of FCA). The categories include 10 natural kind categories[9]: “fruit” (30 exemplars); “vegetables” (30); “professions” (30); “sports” (30); the animal categories “amphibians” (5), “birds” (30), “fish” (23), “insects” (26), “mammals” (30), and “reptiles” (22). In addition, they include 6 artifact categories: “clothing” (29 exemplars), “kitchen utensils” (33), “musical instruments” (27), “tools” (30), “vehicles” (30), and “weapons” (20). The exemplars are selected to be representative of the categories; for instance, the animal categories cover a rather large part of the known animal domain.

The Dutch data also contains information on features (attributes in the sense of FCA). Both objects (exemplars) and attributes (features) were obtained by processes described in [11]. For the obtained objects and attributes, data describing which objects have which attributes was also obtained. Consequently, various matrices (called exemplar by feature applicability matrices by the authors) describing which objects have which attributes were obtained. From the FCA viewpoint, these particular matrices represent particular formal contexts $\langle X, Y, I\rangle$ in a straightforward manner (X and Y are the sets of exemplars and features covered by the matrix and I represents which exemplars have which features).

It is to be noted that two ways of obtaining attributes were used in [11]. Respondents were either asked to list attributes for a given category (these are called category attributes) or for a given exemplar (exemplar attributes). These two kinds of attributes are distinct (for example: category features obtained for the category “fish” overlap with the union of exemplar features obtained for the particular exemplars in this category). As a result, one obtains two versions of the applicability matrices: exemplar by category-feature matrices, and exemplar by exemplar-feature matrices.

Each applicability matrix had been filled out separately by four participants in the study (i.e. the participants were filling out whether objects have attributes). To obtain a single matrix out of these four, we required at least two participants to agree. That is, we defined the corresponding formal context $\langle X, Y, I\rangle$ as follows:

$$\langle x, y\rangle \in I \quad \text{iff at least 2 participants claim that } x \text{ has } y$$

In our study, we utilize four of the formal contexts corresponding to the matrices described in the previous paragraphs. These are described by the following table:

dataset	objects	attributes	density
AnimalCategory	129	225	0.32
AnimalExemplar	129	764	0.13
ArtifactCategory	166	301	0.23
ArtifactExemplar	166	1295	0.09

The table describes the numbers of objects and attributes, and the density of the formal context (e.g., 0.32 means that 32% of the 129×225 entries in the AnimalCategory matrix have value 1).

[9] In this list, we use plural in category names, as the authors do; below, we use singular, i.e. “bird” rather than “birds” to be consistent with our previous writings.

Table 1
Typicality ratings for "bird"; AnimalCategory data.

typ_{SMC} order	typ_{SMC}	typ_{J} order	typ_{J}	typ_{rm} order	typ_{rm}	typ_{HJ} order	typ_{HJ}
woodpecker	0.920	woodpecker	0.783	owl	0.295	sparrow	19.179
blackbird	0.917	blackbird	0.780	parrot	0.295	blackbird	18.821
magpie	0.916	cuckoo	0.777	falcon	0.293	robin	18.321
cuckoo	0.916	magpie	0.774	duck	0.292	dove	18.143
robin	0.913	robin	0.770	dove	0.291	crow	18.107
swallow	0.911	swallow	0.766	sparrow	0.288	seagull	17.964
crow	0.910	crow	0.763	parakeet	0.287	canary	17.893
peacock	0.908	chickadee	0.762	swallow	0.287	magpie	17.893
chickadee	0.908	sparrow	0.759	cuckoo	0.285	swallow	17.857
seagull	0.907	falcon	0.757	chickadee	0.285	parakeet	17.643
falcon	0.905	seagull	0.755	blackbird	0.285	chickadee	17.107
sparrow	0.905	owl	0.751	seagull	0.283	eagle	16.926
pheasant	0.904	peacock	0.746	crow	0.282	woodpecker	16.429
pelican	0.902	dove	0.742	woodpecker	0.282	heron	16.107
heron	0.902	parrot	0.741	rooster	0.282	cuckoo	16.000
owl	0.901	pelican	0.739	robin	0.281	owl	16.000
stork	0.901	pheasant	0.737	canary	0.278	parrot	15.857
dove	0.898	canary	0.736	magpie	0.278	falcon	15.500
canary	0.898	parakeet	0.735	chicken	0.276	stork	15.393
parrot	0.896	stork	0.732	pelican	0.275	vulture	15.143
parakeet	0.895	heron	0.729	eagle	0.274	pheasant	13.714
chicken	0.893	chicken	0.724	vulture	0.268	swan	12.821
turkey	0.893	duck	0.721	turkey	0.267	duck	12.786
duck	0.886	turkey	0.718	peacock	0.267	pelican	12.571
rooster	0.884	rooster	0.711	stork	0.266	peacock	12.286
swan	0.882	eagle	0.700	ostrich	0.266	turkey	11.679
eagle	0.881	swan	0.696	swan	0.266	chicken	11.571
ostrich	0.879	ostrich	0.689	pheasant	0.263	ostrich	11.214
vulture	0.865	vulture	0.669	heron	0.258	rooster	11.071
penguin	0.861	penguin	0.653	penguin	0.257	penguin	8.643

An important question is whether the 16 categories, around which the Dutch data is developed and which are represented by sets of exemplars as described above, actually represent formal concepts; that is, whether the sets of exemplars actually form extents of formal concepts in the formal contexts obtained from the considered applicability matrices.[10] Interestingly, most of the 16 categories indeed represent formal concepts. In particular, this is true of the categories "bird," "fish," and "mammal" which we examine in detail below.[11]

Obtained degrees of typicality

For each of the three concepts mentioned above, we computed the degrees of typicality typ_{SMC}, typ_{J}, and typ_{rm} for all objects in the extent of the concept. We present the results for the AnimalCategory data; for the AnimalExemplar data, our observations are similar.[12] The results are shown in Tables 1 ("bird"), 2 ("fish"), and 3 ("mammal"). In addition to the three computed typicalities, the tables also display the typicality degrees obtained by humans, which are part of the Dutch data and which we denote by typ_{HJ}. Note that we keep the values of typ_{HJ} as they are stored in the Dutch data: They range between 1 and 20 since they are obtained as average degrees assigned by respondents on a twenty-element scale. Each table therefore contains four lists of object-typicality pairs, corresponding to the four kinds of typicality (typ_{SMC}, typ_{J}, typ_{rm}, and typ_{HJ}), and each list is ordered by degrees of typicality. The typicality data displayed in the three tables is also displayed in Figs. 1, 2, and 3.

Thus, for instance, the last two columns of Table 1 display a list of bird exemplars sorted by typicality obtained by human judgment along with the typicality degrees: sparrow is the most typical bird by human judgment, followed by blackbird, robin, etc. On the other hand, penguin, rooster and ostrich are considered the least typical of the available exemplars. Intuitively, this sorted list makes sense. By and large, each of the three other lists of exemplars makes intuitively sense as well. One also observes, for the most part, agreement of each of these three lists, which are obtained by our formulas for computing typicality, with the list obtained by human judgment: The birds typical by human judgment generally tend to be typical according to typ_{SMC}, typ_{J}, and typ_{rm}, and the same may be said of untypical birds.

[10] Note that a given a set A is an extent in a given formal context $\langle X, Y, I\rangle$ iff $A = A^{\uparrow\downarrow}$, i.e. the test is straightforward.

[11] Selection of concepts for a detailed exposition is due to lack of space. Our observations below are representative for what we were able to observe for the other concepts and data.

[12] Recall that AnimalCategory and AnimalExemplar have the same sets of objects but differ in their sets of attributes. While the three categories represent formal concepts in both datasets, the computed typicality degrees are different for the two datasets as a result of the difference in attribute sets, because our formulas for typicality depend on the attribute sets.

Table 2
Typicality ratings for “fish”; AnimalCategory data.

typ_{SMC} order	typ_{SMC}	typ_{J} order	typ_{J}	typ_{rm} order	typ_{rm}	typ_{HJ} order	typ_{HJ}
stickleback	0.927	stickleback	0.813	eel	0.291	goldfish	18.893
plaice	0.925	plaice	0.808	salmon	0.289	salmon	18.393
sardine	0.922	sardine	0.804	pike	0.287	cod	18.107
cod	0.921	cod	0.798	stickleback	0.286	trout	17.893
swordfish	0.920	sole	0.797	sardine	0.285	herring	17.071
sole	0.920	carp	0.795	carp	0.284	pike	16.286
pike	0.920	pike	0.795	plaice	0.284	carp	16.000
carp	0.919	trout	0.789	piranha	0.282	plaice	15.962
trout	0.916	salmon	0.789	herring	0.282	eel	15.679
ray	0.915	swordfish	0.788	flatfish	0.282	sardine	15.536
salmon	0.915	herring	0.785	ray	0.280	piranha	15.321
herring	0.915	flatfish	0.783	trout	0.280	sole	15.143
flatfish	0.915	eel	0.780	sole	0.280	stickleback	14.750
eel	0.911	ray	0.779	cod	0.278	swordfish	14.643
anchovy	0.908	anchovy	0.764	anchovy	0.276	ray	14.500
piranha	0.880	piranha	0.708	swordfish	0.275	flatfish	14.321
goldfish	0.878	goldfish	0.687	goldfish	0.259	shark	13.214
squid	0.868	squid	0.661	squid	0.252	anchovy	13.143
shark	0.847	shark	0.624	shark	0.250	squid	10.679
sperm whale	0.840	sperm whale	0.601	sperm whale	0.235	whale	10.429
dolphin	0.812	dolphin	0.561	dolphin	0.231	sperm whale	9.893
whale	0.804	whale	0.552	whale	0.231	orca	9.857
orca	0.803	orca	0.550	orca	0.230	dolphin	9.179

Table 3
Typicality ratings for “mammal”; AnimalCategory data.

typ_{SMC} order	typ_{SMC}	typ_{J} order	typ_{J}	typ_{rm} order	typ_{rm}	typ_{HJ} order	typ_{HJ}
zebra	0.917	zebra	0.754	dog	0.265	cat	18.536
llama	0.914	kangaroo	0.747	cat	0.262	dog	18.536
kangaroo	0.914	llama	0.741	monkey	0.261	monkey	17.929
dromedary	0.911	dromedary	0.738	horse	0.259	lion	17.679
deer	0.908	deer	0.737	lion	0.258	cow	17.607
donkey	0.907	giraffe	0.726	squirrel	0.255	horse	17.536
giraffe	0.906	donkey	0.725	mouse	0.255	sheep	17.429
bison	0.898	horse	0.722	tiger	0.253	pig	17.179
squirrel	0.898	squirrel	0.718	rabbit	0.253	tiger	17.071
horse	0.898	bison	0.714	deer	0.251	wolf	17.036
fox	0.896	fox	0.708	wolf	0.249	donkey	16.821
cow	0.896	cow	0.705	fox	0.249	rabbit	16.643
sheep	0.894	monkey	0.702	bison	0.248	deer	16.536
beaver	0.891	lion	0.699	beaver	0.247	elephant	16.250
hamster	0.889	beaver	0.697	elephant	0.245	fox	16.250
elephant	0.888	sheep	0.697	kangaroo	0.245	zebra	16.036
monkey	0.888	elephant	0.691	zebra	0.244	giraffe	15.964
lion	0.887	wolf	0.691	hamster	0.241	mouse	15.679
rhinoceros	0.886	hamster	0.689	dromedary	0.241	rhinoceros	15.143
wolf	0.886	cat	0.683	cow	0.241	polar bear	15.143
cat	0.877	rabbit	0.677	donkey	0.239	bison	15.143
pig	0.877	mouse	0.676	giraffe	0.239	kangaroo	14.750
rabbit	0.877	tiger	0.675	llama	0.238	llama	14.643
mouse	0.876	rhinoceros	0.673	hedgehog	0.237	hippopotamus	14.607
tiger	0.876	dog	0.664	sheep	0.237	hamster	14.571
hedgehog	0.875	hedgehog	0.658	polar bear	0.232	squirrel	14.571
hippopotamus	0.874	pig	0.656	hippopotamus	0.227	dromedary	14.429
polar bear	0.874	polar bear	0.654	pig	0.227	beaver	14.000
dog	0.865	hippopotamus	0.651	rhinoceros	0.227	hedgehog	13.179
bat	0.834	bat	0.579	bat	0.225	bat	10.857

General remarks on analyzing typicality data

Note at this point two important aspects that need to be kept in mind in our examination. One concerns human judgment on typicality and is known from the literature: Even though human judgment scores are sometimes called the ground truth, the scores may hardly be regarded as objective. Namely, typicality is subjective to a certain extent as it depends on the experience of the respondent, cultural background and other factors. Secondly, since the computed typicalities rely on the available attributes, the attributes need to describe the objects well: they need to describe the domain of inquiry in a

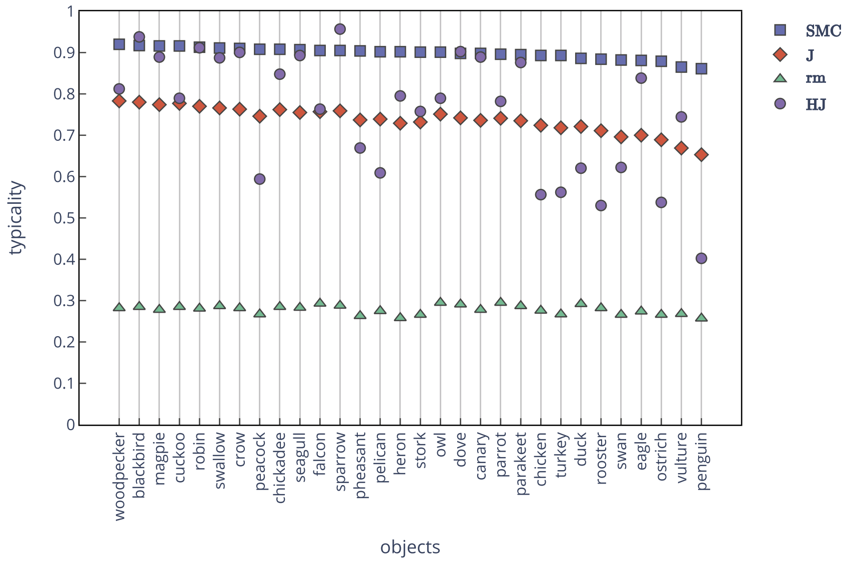

Fig. 1. Typicality ratings for "bird" with objects ordered by values of typ_{SMC} and the values of typ_{HJ} rescaled to $[0, 1]$; AnimalCategory data.

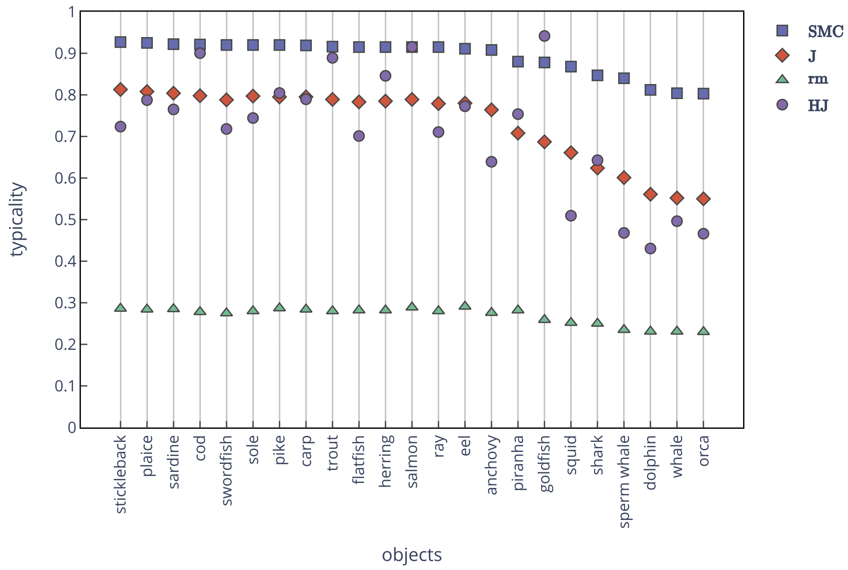

Fig. 2. Typicality ratings for "fish" with objects ordered by values of typ_{SMC} and the values of typ_{HJ} rescaled to $[0, 1]$; AnimalCategory data.

sufficiently informative and balanced way.[13] That is, there need to be enough attributes, describing relevant aspects of the domain, and the attributes must not be redundant (otherwise, the aspect described by redundant attributes would obtain an inappropriately large weight). It is for these reasons that one may hardly expect complete or nearly complete agreement of the computed typicalities with human judgment. As we demonstrate below, a closer examination nevertheless reveals reasonable agreements.

Three methods for analyzing typicality data

To assess the agreements of the typicalities and the sorted lists based on them in a more precise manner, we used three methods. First, we used the well-known Kendall tau rank correlation coefficients τ_b. Kendall tau measures ordinal

[13] In the context of examination of basic level, this is also observed in [6,7].

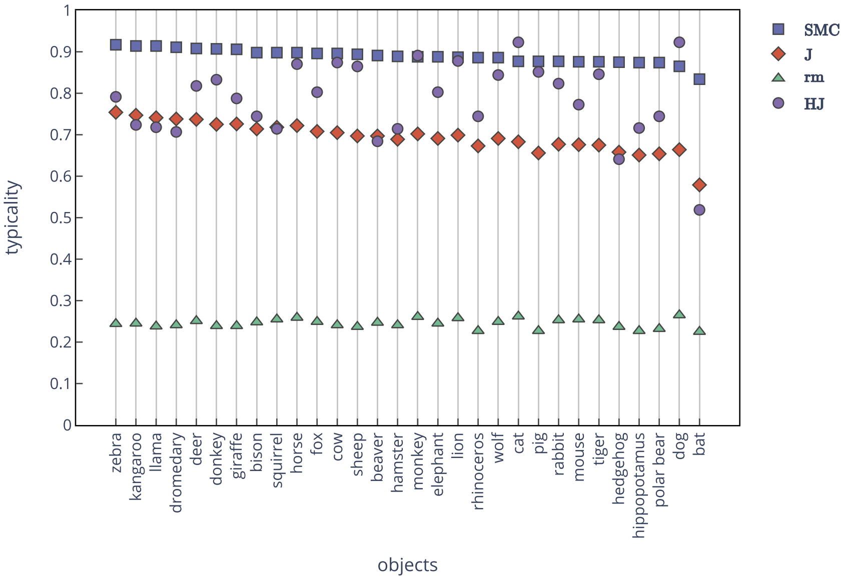

Fig. 3. Typicality ratings for "mammal" with objects ordered by values of typ_{SMC} and the values of typ_{HJ} rescaled to $[0,1]$; AnimalCategory data.

association between two quantities, i.e. between two typicalities in our case. In particular, its value measures to what extent the ordering of exemplars in one list agrees with the ordering of exemplars in the other list. The coefficient ranges from 1 (same ordering) to -1 (inverse, i.e. opposite ordering). We used τ_b to account for ties in typicality values and used its implementation in a Python library [26].

Secondly, we used the $\tilde{\gamma}$ rank correlation coefficient of [8] (called the robust rank correlation coefficient by the authors). Namely, the Kendall tau only takes into account the orderings of exemplars and disregards the typicality degrees on which the ordering is based. One may object to this as follows. Consider three lists, each consisting of two objects x_1 and x_2, along with their typicalities, say

$$l_1 = \langle\langle x_1, 0.85\rangle, \langle x_2, 0.1\rangle\rangle,$$
$$l_2 = \langle\langle x_1, 0.85\rangle, \langle x_2, 0.8\rangle\rangle, \text{ and}$$
$$l_3 = \langle\langle x_2, 0.9\rangle, \langle x_1, 0.8\rangle\rangle.$$

The Kendall τ_b of l_1 and l_3 is -1 (opposite ordering), which is the same as τ_b of l_2 and l_3 (opposite ordering as well), since only the orderings matter. However, since we naturally also take the typicality degrees into account, l_2 and l_3 are much better correlated (since the typicality degrees are very close) than l_1 and l_3. The $\tilde{\gamma}$ coefficient resolves this by taking closeness of degrees into account. In particular, we set the parameter r, which controls what the method considers as close values of typicality, to $r = 0.2$.

Thirdly, we employed the idea put forward in our previous study [7] to alleviate the strictness of rank correlation consisting in basically looking solely at agreement of two compared orderings of objects. One might argue that rather than to examine agreement in ordering, it is more interesting to examine whether the set of the top r objects (i.e., r most typical objects) in one list is similar to the set of top r objects in the other list for various values of r. For a given typicality assignment M (e.g. M = typ_{J}), we denote the set of the top r objects in the list corresponding to M as

$$Top_r^{\mathrm{M}}.$$

For this, we assume that (a) if the $(r+1)$-st, ..., $(r+k)$-th objects are tied with the r-th one, i.e. have the same value of typicality, we add these k objects to Top_r^{M}; (b) we do not include objects with typicality equal to 0. Now, given typicality assignments M and N, we are interested in whether and to what extent are the sets Top_r^{M} and Top_r^{N} similar. For this purpose, we proceed as follows. For objects x_1 and x_2, we denote by $s(x_1, x_2)$ a suitable defined similarity degree (a number in $[0,1]$ in our case).

Below, we use

$$s(x_1, x_2) = sim_{\mathrm{J}}(\{x_1\}^{\uparrow}, \{x_2\}^{\uparrow}),$$

i.e. $s(x_1, x_2)$ equals the Jaccard index of the sets $\{x_1\}^{\uparrow}$ and $\{x_2\}^{\uparrow}$ of attributes for objects x_1 and x_2, respectively; cf. (4).

Table 4
Correlations of typicalities; AnimalCategory and AnimalExemplar data.

dataset	concept	τ_b correlation				$\tilde{\gamma}$ correlation			
AnimalCategory	bird		typ_{J}	typ_{rm}	typ_{HJ}		typ_{J}	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.862	0.196	0.445	typ_{SMC}	0.978	0.426	0.634
		typ_{J}		0.334	0.5	typ_{J}		0.601	0.691
		typ_{rm}			0.319	typ_{rm}			0.496
	fish		typ_{J}	typ_{rm}	typ_{HJ}		typ_{J}	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.919	0.531	0.412	typ_{SMC}	0.999	0.932	0.776
		typ_{J}		0.581	0.462	typ_{J}		0.93	0.758
		typ_{rm}			0.47	typ_{rm}			0.784
	mammal		typ_{J}	typ_{rm}	typ_{HJ}		typ_{J}	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.871	0.014	-0.053	typ_{SMC}	0.981	0.042	-0.005
		typ_{J}		0.133	0.002	typ_{J}		0.28	0.115
		typ_{rm}			0.413	typ_{rm}			0.547
AnimalExemplar	bird		typ_{J}	typ_{rm}	typ_{HJ}		typ_{J}	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.839	0.269	0.454	typ_{SMC}	0.968	0.398	0.652
		typ_{J}		0.43	0.56	typ_{J}		0.61	0.784
		typ_{rm}			0.505	typ_{rm}			0.717
	fish		typ_{J}	typ_{rm}	typ_{HJ}		typ_{J}	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.927	0.428	0.333	typ_{SMC}	0.995	0.736	0.615
		typ_{J}		0.47	0.32	typ_{J}		0.749	0.553
		typ_{rm}			0.249	typ_{rm}			0.259
	mammal		typ_{J}	typ_{rm}	typ_{HJ}		typ_{J}	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.582	-0.241	-0.345	typ_{SMC}	0.856	-0.449	-0.472
		typ_{J}		0.177	-0.002	typ_{J}		0.274	-0.025
		typ_{rm}			0.595	typ_{rm}			0.729

Finally, for two typicality assignments, M and N, and a given $r = 1, 2, 3, \ldots$, we define

$$S(Top_r^{\mathrm{M}}, Top_r^{\mathrm{N}}) = \min(I_{\mathrm{MN}}, I_{\mathrm{NM}})$$

where

$$I_{\mathrm{MN}} = \frac{\sum_{x_1 \in Top_r^{\mathrm{M}}} \max_{x_2 \in Top_r^{\mathrm{N}}} s(x_1, x_2)}{|Top_r^{\mathrm{M}}|}$$

and, symmetrically,

$$I_{\mathrm{NM}} = \frac{\sum_{x_2 \in Top_r^{\mathrm{N}}} \max_{x_1 \in Top_r^{\mathrm{M}}} s(x_1, x_2)}{|Top_r^{\mathrm{N}}|}.$$

According to basic rules of fuzzy logic, $S(Top_r^{\mathrm{M}}, Top_r^{\mathrm{N}})$ may naturally be interpreted as the truth degree of the proposition "for most objects in Top_r^{M} there is a similar object in Top_r^{N} and vice versa." Due to this interpretation and since S is actually a reflexive and symmetric fuzzy relation [3,28], S is a good candidate for measuring similarity of sets of objects. High values of S indicate high similarity and $S(Top_r^{\mathrm{M}}, Top_r^{\mathrm{N}}) = 1$ takes place if and only if $Top_r^{\mathrm{M}} = Top_r^{\mathrm{N}}$.

Results of analyses

Consider first the rank correlations τ_b and $\tilde{\gamma}$, which are shown for the concepts "bird," "fish," and "mammal" in Table 4. The table presents, both for the AnimalCategory and AnimalExemplar data, and for the three concepts in this data all the six correlation coefficients τ_b and six correlation coefficients $\tilde{\gamma}$ for the four observed typicalities typ_{SMC}, typ_{J}, typ_{rm}, and typ_{HJ}.

Let us first examine the τ_b correlations of the three computed typicalities with human judgment. Note first that according to a commonly accepted interpretation, the values of τ_b may be interpreted as follows: $\tau_b \geq 0.3$, $0.2 \leq \tau_b < 0.3$, $0.1 \leq \tau_b < 0.2$, and $0.0 \leq \tau_b < 0.1$ indicate strong, moderate, weak and very weak correlation, respectively (analogously for negative values). All the computed typicalities, typ_{SMC}, typ_{J}, typ_{rm} display a strong correlation with human judgment in virtually all cases except for the concept "mammal". For this concept, only typ_{rm} exhibits a strong correlation. We only have a partial explanation for this. Namely, we believe that "mammal" is a somewhat problematic concept as regards human judgment of typicality (we observed this when collecting human rankings for "mammal" in the Zoo data; see the next section); we encountered similar difficulties with some other concepts, e.g. "kitchen utensils" (what is a typical kitchen utensil?). Why some concepts are problematic in this sense is a question that should be explored in the future, possibly with the help of psychologists.

Next, let us examine mutual correlations of the three computed typicalities. The data indicates a very strong correlation between typ_{SMC} and typ_{J} in most cases. Furthermore, we see significantly less strong correlations between typ_{SMC} and typ_{rm}, between typ_{J} and typ_{rm}; yet these correlations range from moderate to strong in most cases, except for the above-discussed "mammal".

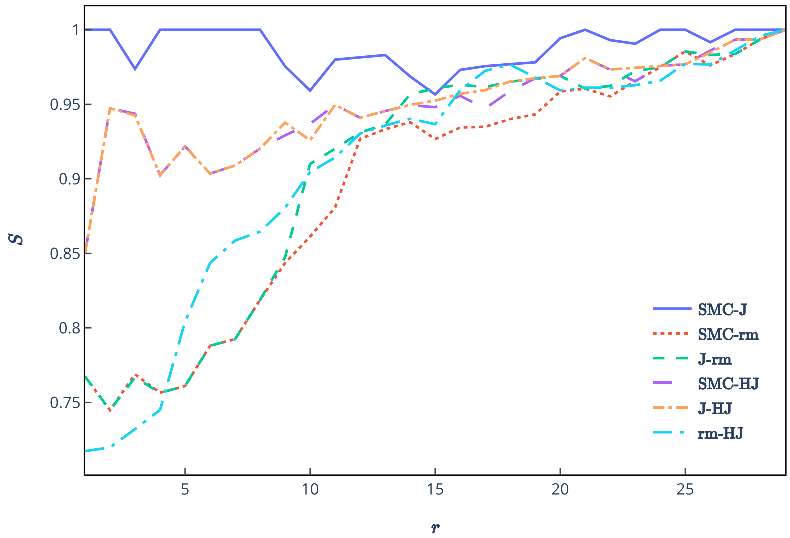

Fig. 4. Similarity S of top r typical objects; concept “bird” in AnimalCategory data. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Similar pattern may be observed for the $\tilde{\gamma}$ correlations. Since we observed by and large the same behavior for all the data and concepts we examined, including the Zoo data presented in the next section, the correlation analysis suggests as interesting the problem to analyze, both experimentally and theoretically, the relationships of the three computed typicalities and, in particular, to focus on why typ_{rm} seems to perform differently from the two rather correlated ones.

Let us next consider our third method of comparison. The mutual similarities of the sets of top r typical objects for the AnimalCategory and the AnimalExemplar data for our three examined concepts are shown in Figs. 4–9. Each figure displays six graphs representing the six mutual similarities $S = S(Top_r^{\mathrm{M}}, Top_r^{\mathrm{N}})$ of the sets of top r typical objects according to typicality M and typicality N (vertical axis), for increasing $r = 1, 2, \ldots$ (horizontal axis). The graphs reveal a similar pattern of relationships we observed with the correlations. For instance, in Fig. 4, the blue graph labeled SMC-J representing the similarities for M = $\mathrm{typ_{SMC}}$ and N = $\mathrm{typ_J}$ shows a high similarity of the sets of top r $\mathrm{typ_{SMC}}$-typical and top r $\mathrm{typ_J}$-typical birds, which confirms—from a different perspective—the very strong rank correlations of these two typicalities observed above. The two lines, one for M = $\mathrm{typ_{SMC}}$ and N = $\mathrm{typ_{HJ}}$, the other for M = $\mathrm{typ_J}$ and N = $\mathrm{typ_{HJ}}$, which also attain high values even for small r confirm strong correlations of these two pairs of typicalities observed above. Naturally, the similarities increase with increasing r which needs to be taken into account when interpreting the graphs.

4.2. Experiments with Zoo data

Zoo data

Zoo is a commonly known dataset [12] and its concepts are mostly well interpretable. It describes 101 animals (objects) by their 17 attributes and has the density of 0.36. We removed the somewhat disputable object “girl” from the data; we renamed one of the two objects denoted “frog” to “frog venomous.” All of the attributes are yes/no attributes except for the attribute describing the number of legs, which we nominally scaled, and an attribute determining the type of animal, which we removed. The scaled data is presented in Table 5 (to save space, objects with the same attributes are put on the same row).

The Zoo data (i.e. the formal context corresponding to the data) contains several formal concepts, among them three concepts that may be interpreted as “bird,” “fish,” and “mammal”. Since concepts with the same interpretation (in different data, however, and thus represented by different sets of objects and attributes) were used in the previous section, we examine typicalities for these concepts.

To be able to perform similar analyses to those we described the previous section, we first obtained human typicality ratings for the objects of the three concepts by means of a questionnaire; see [4] for the data and appendix for the questionnaire. Since the obtained data may be useful for further studies, we describe it to a certain detail. Altogether, 242 respondents participated in the survey. We first split the respondents in four groups (students at Palacky University Olomouc, our coworkers, relatives, and others), since we assumed possibly different reliability of these groups. However, as a correlation analysis revealed high correlations between the average ratings in these groups (Kendall τ_b always higher than 0.6), we merged the groups into a single group for which we computed average typicality ratings. Note also that approximately 56% (136) women and 44% (106) men were among the participants. The median, minimum and maximum age of participants was 23, 17 and 81. The three concepts along with the obtained human judgment of typicality and the three

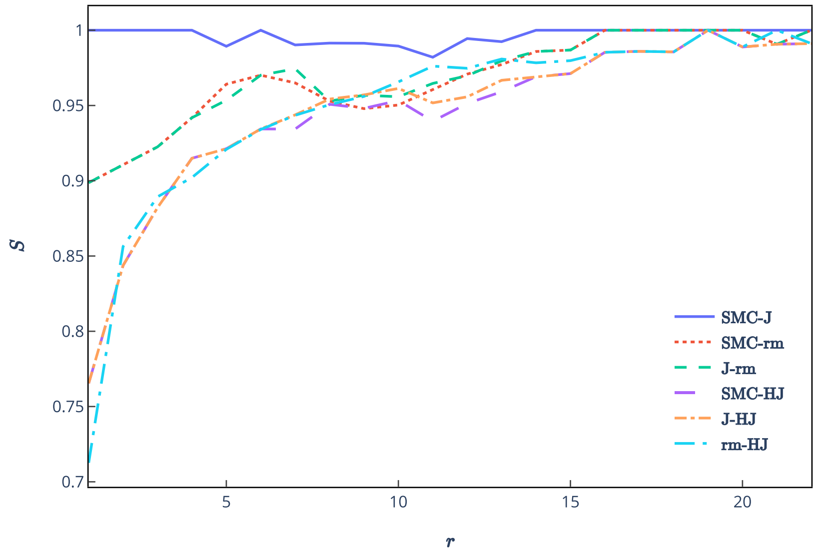

Fig. 5. Similarity S of top r typical objects; concept “fish” in AnimalCategory data.

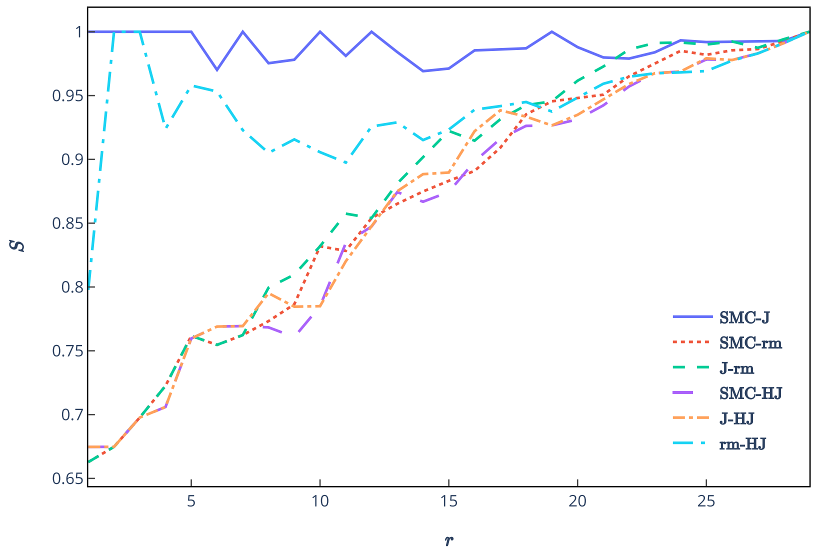

Fig. 6. Similarity S of top r typical objects; concept “mammal” in AnimalCategory data.

computed typicality ratings are shown in Tables 6, 7, and 8, respectively, in the same manner as with the Dutch data. As in the previous section, the typicality data is also displayed in Figs. 10, 11, and 12.

Results of analyses

The correlations of the typicalities for the three concepts are shown in Table 9. The mutual similarities of the set of top r objects by the observed typicalities are displayed in Figs. 13, 14, and 15.

Basically, a similar pattern as for the Dutch data may be observed with the following differences. First, correlations of the computed typicalities to human judgment are generally somewhat smaller, which is due to the fact that the attributes in the Zoo data are considerably less informative compared to the attributes in Dutch data (several animals have the same attributes in the Zoo data but, at the same time, are rather different as regards typicality). This illustrates the need of informative attributes for the computed typicalities to work reasonably well, as mentioned above. In addition, typ_{rm}, which again behaves somewhat differently compared to the strongly correlated typ_{SMC} and typ_{J}, achieves smaller correlation with hu-

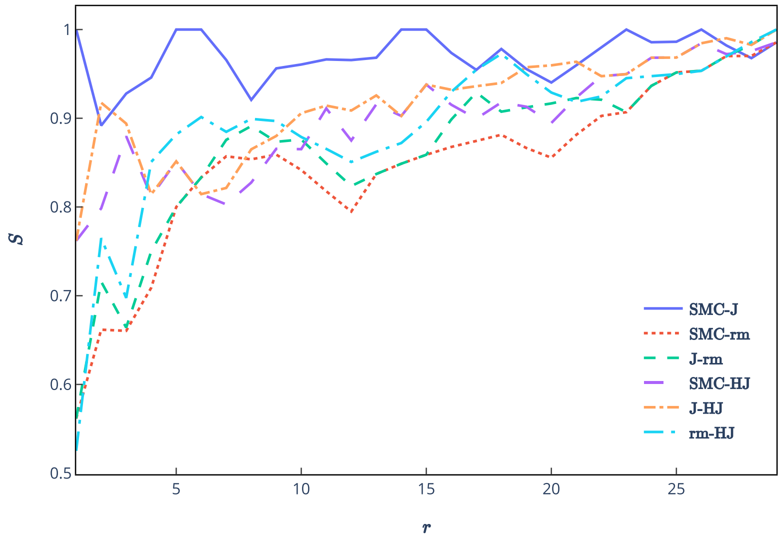

Fig. 7. Similarity S of top r typical objects; concept "bird" in AnimalExemplar data.

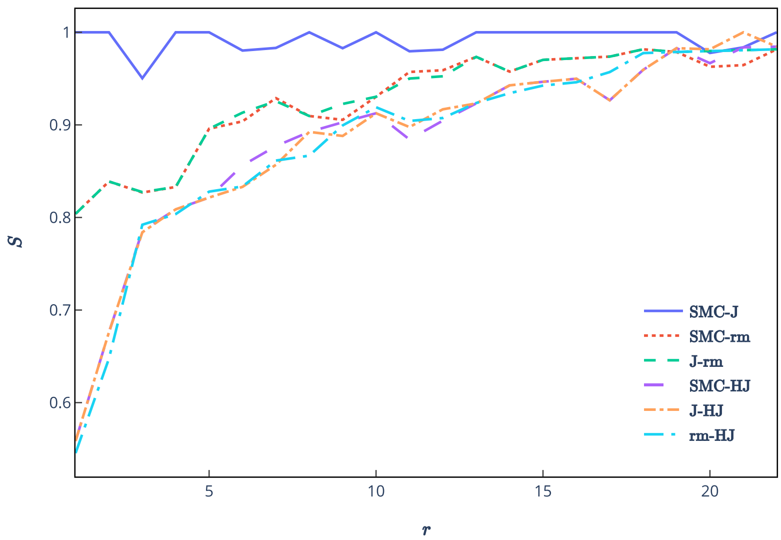

Fig. 8. Similarity S of top r typical objects; concept "fish" in AnimalExemplar data.

man judgment compared to Dutch data. This again calls for a closer examination of the relationship between the computed typicalities.

Finally, let us mention an instructive observation regarding the concept "fish." Here, carp appears to be a problematic exemplar. While most of our respondents consider it the most typical fish, it is considered atypical according to our formulas. The reason is, on the one hand, that the presence of the attribute domestic and absence of predator make carp atypical according to the typicality formulas. On the other hand, the perception of carp by respondents is influenced by other factors, including cultural background, which—as in this case—may be considerably more significant for human judgment than the actual attributes present in the data when it comes to determination of typicality. This phenomenon is discussed in the literature [20] and calls for a closer examination from a general perspective. Note also that the presence of this single problematic exemplar makes the correlations of the computed typicalities with human judgment very weak. When carp is removed from the extent of the concept "fish" (in which case the set of objects no longer forms an extent of a formal concept), the rank correlations with human judgment become considerably stronger; see the correlations in the part "fish (without carp)".

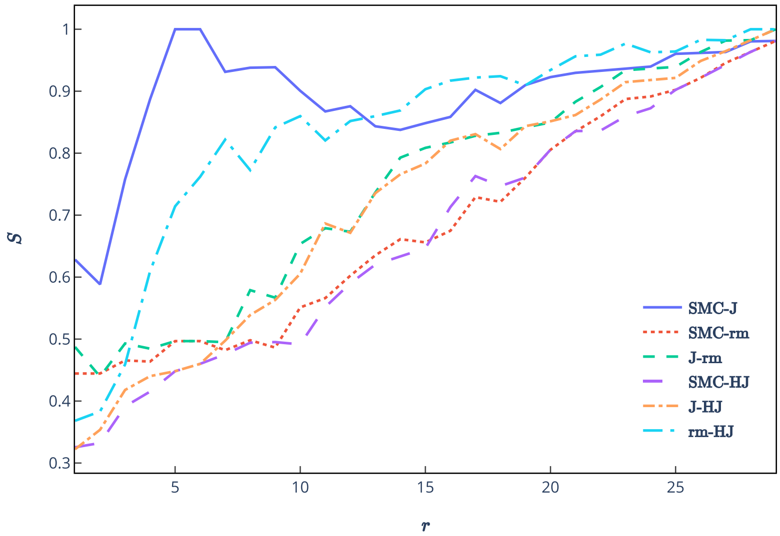

Fig. 9. Similarity S of top r typical objects; concept "mammal" in AnimalExemplar data.

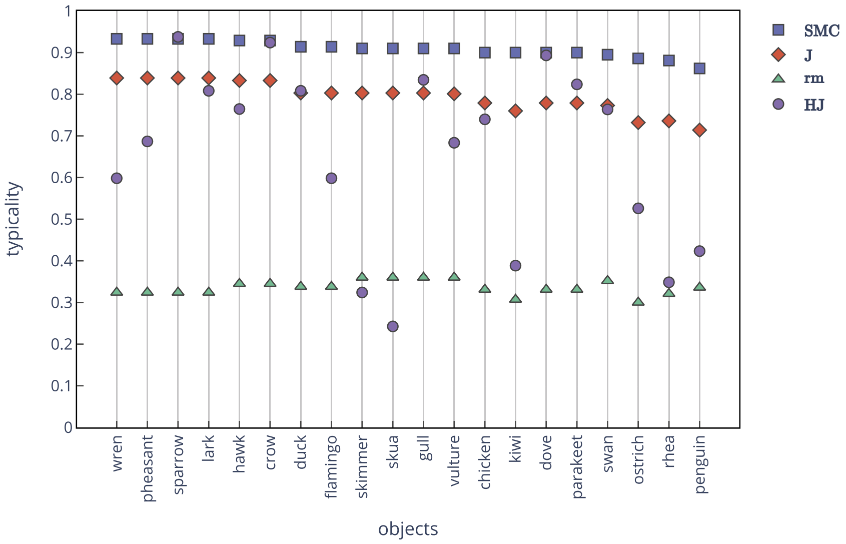

Fig. 10. Typicality ratings for "bird" with objects ordered by values of typ_{SMC} and the values of typ_{HJ} rescaled to $[0, 1]$; Zoo data.

5. Conclusions and further topics

This paper is intended to make first steps in studying and exploiting typicality within a more formalized setting compared to what is common in the literature on the psychology of concepts. The basic aim is to enable a more precise analysis of typicality, both experimental and theoretical.

We proposed a formal definition of typicality, which translates to a general scheme for a formula to compute typicality of objects of a given concept in a given data consisting of objects, attributes and an incidence relation between objects and attributes. Our scheme is based on a basic psychological view of typicality due to Rosch and Mervis. We considered three typicality functions resulting from the general scheme, namely typ_{SMC}, typ_{J}, and typ_{rm}, the last of which was proved equivalent to a function actually proposed by Rosch and Mervis. Experiments performed with the Dutch and the Zoo data revealed that for most concepts in this data for which human judgment on typicality is available, there is a strong agreement between the computed typicalities and human judgment. This finding was confirmed by three kinds of analyses. We also observed a very strong correlation of the functions typ_{SMC} and typ_{J}, and a considerably weaker correlation of these functions

Table 5
Scaled Zoo data.

	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	tail	domestic	catsize	legs 0	legs 2	legs 4	legs 5	legs 6	legs 8
scorpion							×			×	×		×								×
seasnake						×	×	×	×		×		×			×					
dolphin, porpoise				×		×	×	×	×	×		×	×		×	×					
flea, termite			×							×										×	
slug, worm			×							×						×					
tortoise			×						×	×			×		×			×			
clam			×				×									×					
tuatara			×				×	×	×	×			×					×			
slowworm			×				×	×	×	×			×			×					
pitviper			×				×	×	×	×	×		×			×					
haddock, seahorse, sole			×			×		×	×			×	×			×					
carp			×			×		×	×			×	×	×		×					
toad			×			×		×	×	×								×			
crayfish, lobster			×			×	×													×	
starfish			×			×	×												×		
crab			×			×	×											×			
octopus			×			×	×								×						×
seawasp			×			×	×				×					×					
bass, catfish, chub, herring, piranha			×			×	×	×	×			×	×			×					
dogfish, pike, tuna			×			×	×	×	×			×	×		×	×					
stingray			×			×	×	×	×		×	×	×		×	×					
frog			×			×	×	×	×	×								×			
newt			×			×	×	×	×	×			×					×			
frog venomous			×			×	×	×	×	×	×							×			
gnat			×		×					×										×	
ladybird			×		×		×			×										×	
ostrich		×	×						×	×			×		×		×				
kiwi		×	×				×		×	×			×				×				
rhea		×	×				×		×	×			×		×		×				
penguin		×	×			×	×		×	×			×		×		×				
lark, pheasant, sparrow, wren		×	×		×				×	×			×				×				
flamingo		×	×		×				×	×			×		×		×				
chicken, dove, parakeet		×	×		×				×	×			×	×			×				
crow, hawk		×	×		×		×		×	×			×				×				
vulture		×	×		×		×		×	×			×		×		×				
duck		×	×		×	×			×	×			×				×				
swan		×	×		×	×			×	×			×		×		×				
gull, skimmer, skua		×	×		×	×	×		×	×			×				×				
gorilla	×			×				×	×	×					×		×				
cavy	×			×				×	×	×				×				×			
hare, vole	×			×				×	×	×			×					×			
squirrel	×			×				×	×	×			×				×				
antelope, buffalo, deer, elephant, giraffe, oryx	×			×				×	×	×			×		×			×			
wallaby	×			×				×	×	×			×		×		×				
hamster	×			×				×	×	×			×	×				×			
calf, goat, pony, reindeer	×			×				×	×	×			×	×	×			×			
aardvark, bear	×			×			×	×	×	×					×			×			
mole, opossum	×			×			×	×	×	×			×					×			
pussycat	×			×			×	×	×	×			×	×	×			×			
mink	×			×		×	×	×	×	×			×		×			×			
seal	×			×		×	×	×	×	×		×			×	×					
sealion	×			×		×	×	×	×	×		×	×		×		×				
fruitbat, vampire	×			×	×			×	×	×			×				×				
housefly, moth	×		×		×					×										×	
wasp	×		×		×					×	×									×	
honeybee	×		×		×					×	×			×						×	
platypus	×		×	×		×	×		×	×			×		×			×			
boar, cheetah, leopard, lion, lynx, mongoose	×			×			×	×	×	×			×		×			×			
polecat, puma, raccoon, wolf	×			×			×	×	×	×			×		×			×			

with typ_{rm}, which implies a need for a further detailed analysis of the proposed typicality functions. We also pointed out problems to consider from a psychological viewpoint which are relevant for evaluation of formal definitions of typicality.

Major topics to be explored in the future are the following:

Table 6
Typicality ratings for "bird"; Zoo data.

typ_{SMC} order	typ_{SMC}	typ_J order	typ_J	typ_{rm} order	typ_{rm}	typ_{HJ} order	typ_{HJ}
wren	0.933	wren	0.839	vulture	0.360	sparrow	4.751
pheasant	0.933	pheasant	0.839	gull	0.360	crow	4.696
sparrow	0.933	sparrow	0.839	skua	0.360	dove	4.573
lark	0.933	lark	0.839	skimmer	0.360	gull	4.338
hawk	0.929	hawk	0.833	swan	0.352	parakeet	4.295
crow	0.929	crow	0.833	hawk	0.345	duck	4.233
duck	0.914	duck	0.803	crow	0.345	lark	4.233
flamingo	0.914	flamingo	0.803	duck	0.338	hawk	4.058
vulture	0.910	gull	0.803	flamingo	0.338	swan	4.054
gull	0.910	skua	0.803	penguin	0.336	chicken	3.959
skua	0.910	skimmer	0.803	chicken	0.331	pheasant	3.747
skimmer	0.910	vulture	0.801	parakeet	0.331	vulture	3.734
chicken	0.900	chicken	0.779	dove	0.331	flamingo	3.393
kiwi	0.900	dove	0.779	pheasant	0.324	wren	3.393
dove	0.900	parakeet	0.779	lark	0.324	ostrich	3.104
parakeet	0.900	swan	0.773	sparrow	0.324	penguin	2.693
swan	0.895	kiwi	0.760	wren	0.324	kiwi	2.554
ostrich	0.886	rhea	0.736	rhea	0.321	rhea	2.394
rhea	0.881	ostrich	0.732	kiwi	0.307	skimmer	2.296
penguin	0.862	penguin	0.714	ostrich	0.300	skua	1.970

Table 7
Typicality ratings for "fish"; Zoo data.

typ_{SMC} order	typ_{SMC}	typ_J order	typ_J	typ_{rm} order	typ_{rm}	typ_{HJ} order	typ_{HJ}
bass	0.963	bass	0.913	stingray	0.385	carp	4.860
catfish	0.963	catfish	0.913	dogfish	0.381	pike	4.711
chub	0.963	chub	0.913	pike	0.381	catfish	4.513
herring	0.963	herring	0.913	tuna	0.381	haddock	4.430
piranha	0.963	piranha	0.913	bass	0.366	bass	4.429
dogfish	0.945	dogfish	0.875	catfish	0.366	tuna	4.418
haddock	0.945	pike	0.875	chub	0.366	piranha	4.054
pike	0.945	tuna	0.875	herring	0.366	herring	4.018
seahorse	0.945	haddock	0.868	piranha	0.366	dogfish	3.104
sole	0.945	seahorse	0.868	carp	0.337	chub	2.992
tuna	0.945	sole	0.868	haddock	0.333	stingray	2.948
carp	0.905	stingray	0.803	seahorse	0.333	seahorse	2.242
stingray	0.905	carp	0.788	sole	0.333	sole	2.042

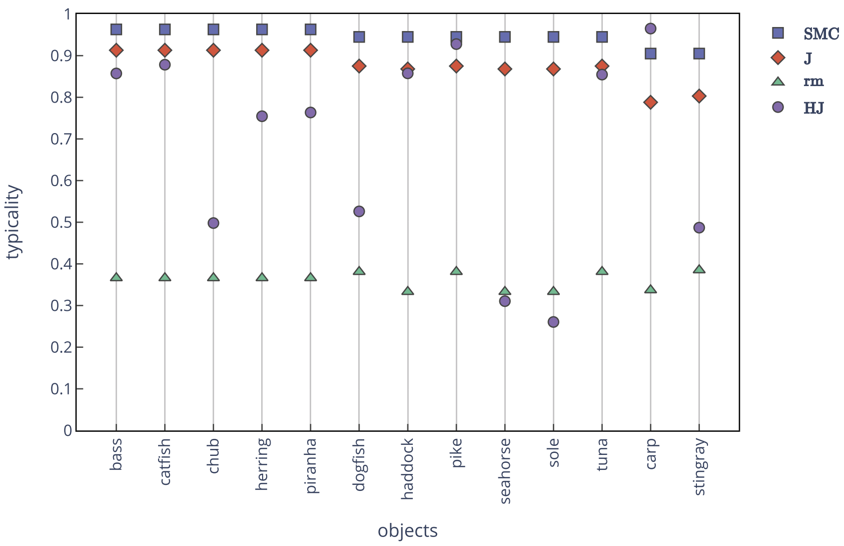

Fig. 11. Typicality ratings for "fish" with objects ordered by values of typ_{SMC} and the values of typ_{HJ} rescaled to [0, 1]; Zoo data.

Table 8
Typicality ratings for "mammal"; Zoo data.

typ_{SMC} order	typ_{SMC}	typ_J order	typ_J	typ_{rm} order	typ_{rm}	typ_{HJ} order	typ_{HJ}
wolf	0.912	wolf	0.822	pussycat	0.383	pussycat	4.714
cheetah	0.912	cheetah	0.822	mink	0.382	calf	4.636
lynx	0.912	lynx	0.822	leopard	0.375	goat	4.521
lion	0.912	mongoose	0.822	raccoon	0.375	bear	4.438
leopard	0.912	polecat	0.822	puma	0.375	gorilla	4.423
polecat	0.912	leopard	0.822	polecat	0.375	wolf	4.397
puma	0.912	lion	0.822	mongoose	0.375	hare	4.328
raccoon	0.912	puma	0.822	lynx	0.375	lion	4.278
boar	0.912	boar	0.822	lion	0.375	deer	4.269
mongoose	0.912	raccoon	0.822	wolf	0.375	pony	4.088
deer	0.910	buffalo	0.810	boar	0.375	boar	4.059
elephant	0.910	elephant	0.810	cheetah	0.375	cheetah	4.033
giraffe	0.910	oryx	0.810	calf	0.358	elephant	4.021
oryx	0.910	giraffe	0.810	goat	0.358	lynx	3.967
buffalo	0.910	antelope	0.810	reindeer	0.358	giraffe	3.942
antelope	0.910	deer	0.810	pony	0.358	leopard	3.934
opossum	0.886	pussycat	0.774	sealion	0.357	squirrel	3.908
mole	0.886	mink	0.765	elephant	0.350	puma	3.900
hare	0.883	pony	0.762	giraffe	0.350	hamster	3.892
vole	0.883	goat	0.762	buffalo	0.350	cavy	3.863
pussycat	0.881	reindeer	0.762	deer	0.350	reindeer	3.801
pony	0.879	calf	0.762	antelope	0.350	antelope	3.791
reindeer	0.879	mole	0.760	oryx	0.350	buffalo	3.770
mink	0.879	opossum	0.760	mole	0.338	wallaby	3.628
goat	0.879	hare	0.746	opossum	0.338	raccoon	3.619
calf	0.879	vole	0.746	platypus	0.337	vole	3.608
bear	0.876	bear	0.741	bear	0.333	polecat	3.404
aardvark	0.876	aardvark	0.741	aardvark	0.333	mink	3.311
hamster	0.852	hamster	0.703	hamster	0.321	mole	3.302
wallaby	0.850	wallaby	0.690	wallaby	0.320	dolphin	3.186
squirrel	0.824	squirrel	0.631	hare	0.313	opossum	3.172
cavy	0.817	sealion	0.622	vole	0.313	oryx	3.016
gorilla	0.814	cavy	0.621	seal	0.312	seal	2.917
platypus	0.788	platypus	0.620	porpoise	0.308	aardvark	2.899
fruitbat	0.781	gorilla	0.610	dolphin	0.308	sealion	2.837
sealion	0.781	fruitbat	0.581	fruitbat	0.286	fruitbat	2.474
vampire	0.781	vampire	0.581	vampire	0.286	mongoose	2.438
seal	0.738	seal	0.549	squirrel	0.283	porpoise	2.425
porpoise	0.731	porpoise	0.540	cavy	0.280	vampire	2.242
dolphin	0.731	dolphin	0.540	gorilla	0.279	platypus	2.230

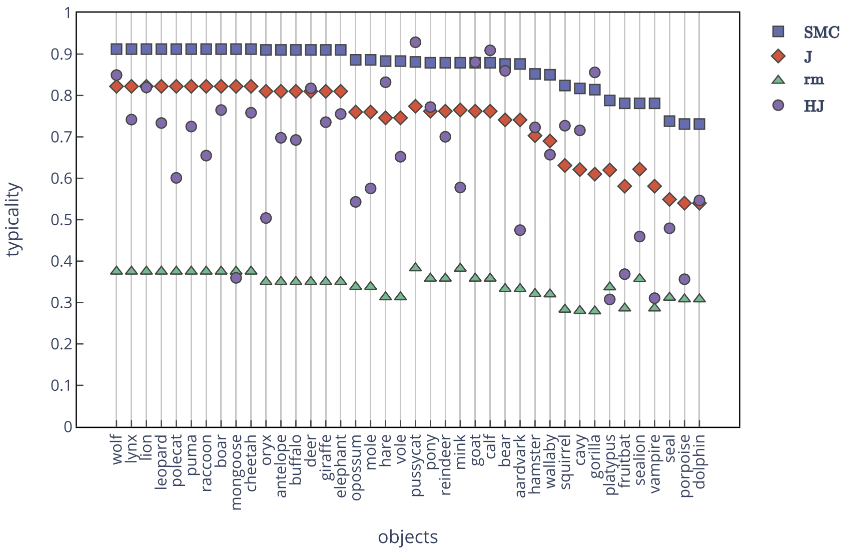

Fig. 12. Typicality ratings for "mammal" with objects ordered by values of typ_{SMC} and the values of typ_{HJ} rescaled to $[0, 1]$; Zoo data.

Table 9
Correlations of typicalities; Zoo data.

dataset	concept	τ_b correlation				$\tilde{\gamma}$ correlation			
Zoo	bird		typ_J	typ_{rm}	typ_{HJ}		typ_J	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.948	0.012	0.267	typ_{SMC}	0.995	0.073	0.45
		typ_J		0.063	0.264	typ_J		0.229	0.468
		typ_{rm}			0.044	typ_{rm}			0.057
	fish		typ_J	typ_{rm}	typ_{HJ}		typ_J	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.916	-0.106	0.0	typ_{SMC}	1.0	0.088	-0.059
		typ_J		0.065	0.058	typ_J		0.273	-0.074
		typ_{rm}			0.029	typ_{rm}			0.132
	fish (without carp)		typ_J	typ_{rm}	typ_{HJ}		typ_J	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.906	-0.243	0.211	typ_{SMC}	1.0	-0.235	0.423
		typ_J		-0.04	0.279	typ_J		0.081	0.418
		typ_{rm}			0.139	typ_{rm}			0.402
	mammal		typ_J	typ_{rm}	typ_{HJ}		typ_J	typ_{rm}	typ_{HJ}
		typ_{SMC}	0.904	0.599	0.238	typ_{SMC}	0.995	0.831	0.556
		typ_J		0.695	0.27	typ_J		0.873	0.482
		typ_{rm}			0.233	typ_{rm}			0.296

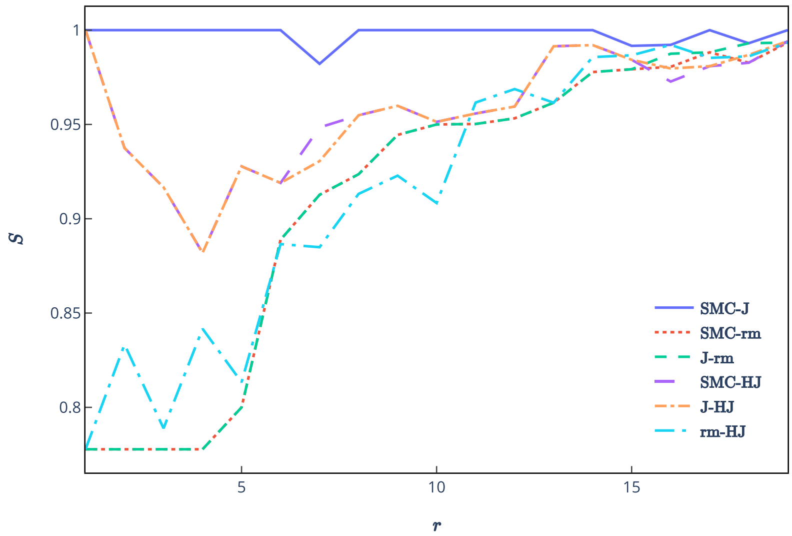

Fig. 13. Similarity S of top r typical objects; concept "bird" in Zoo data.

- Crucial for performing experimental evaluation of formal approaches to typicality and surrounding phenomena is availability of quality data that includes data describing human judgment. Dutch data seems to be the most comprehensive available data for this purpose. Obtaining such data and making the data publicly available appears to be an important goal.
- Our experience with obtaining human judgment of typicality suggests that data describing human judgment should not be taken as "ground truth" (to use a term often mentioned in the psychological literature), for which a perfect fit is required with a given formula for computing typicality. Namely, the data on human judgment may have its own issues some of which are mentioned above. Rather than seeking a perfect fit, an evaluation of a proposed formalization of typicality needs to be performed with caution. This issue seems to point out an important methodological question which involves both psychological and mathematical aspects.
- As mentioned above, our experiments suggest that it is important to analyze, experimentally and theoretically, further relationships between the three proposed typicality functions, as well as to explore further instances of our general scheme for typicality formulas. In particular, explorations of further similarity functions is needed. As an example, we performed experiments with similarity functions which disregard attributes from the intent of the concept for which typicality is evaluated; this approach seems to have some advantages over the three similarity functions described above, such as better distinction between typical and atypical objects.
- Formalization of other existing psychological views of typicality, such as those mentioned in section 2.2, clearly represents a related, important goal. This includes the possibility to take into account the second part of Rosch and Mervis view of typicality, mentioned in section 2.2, which regards similarity to objects in other categories. More radical depar-

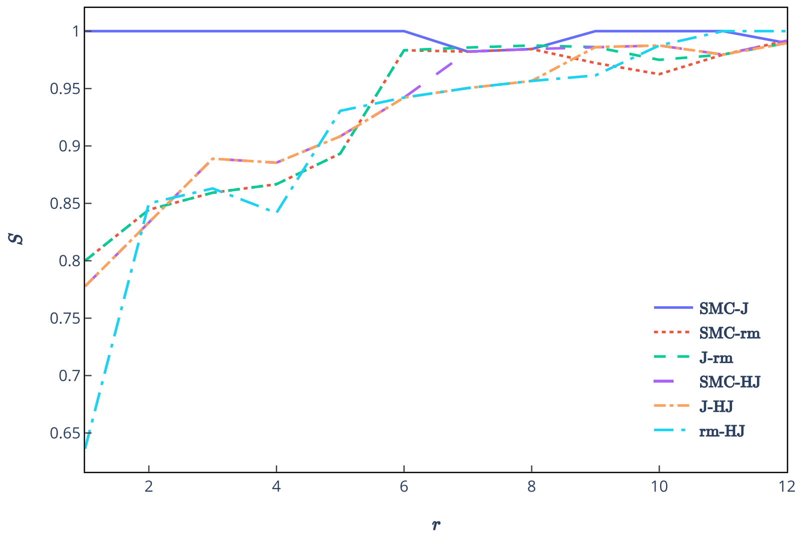

Fig. 14. Similarity S of top r typical objects; concept “fish” in Zoo data.

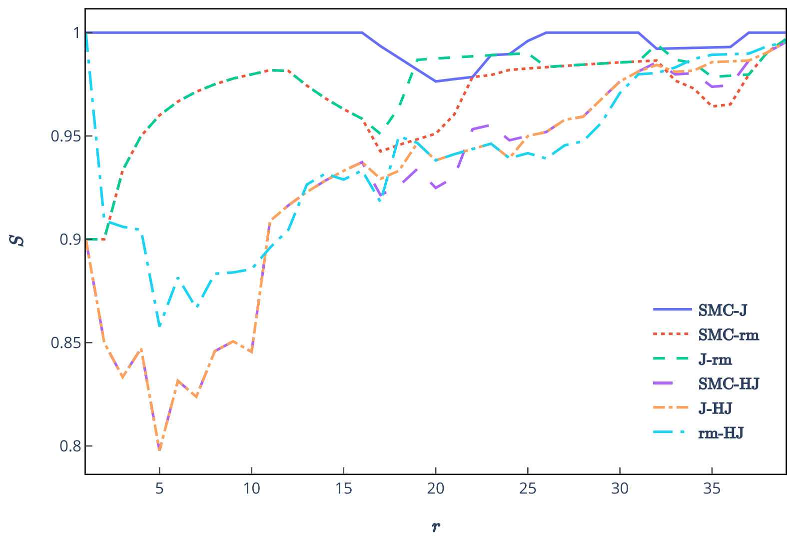

Fig. 15. Similarity S of top r typical objects; concept “mammal” in Zoo data.

tures from our present approach would take dependence of typicality on context into account (in a general sense of the notion of context), as suggested, e.g. in [27] and [10].

- In addition to typicality of objects, typicality of attributes may be explored. At the first sight, this seems just a dual case of typicality of objects. From a psychological point of view, however, typicality of attributes has a rather different role; see e.g. [20]. Methods to determine typicality of attributes shall thus be explored.
- Due to considerable cognitive significance of typicality, it seems natural to exploit typicality in machine learning and data analysis and thus extend the existing attempts mentioned in section 3.2.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This paper is an extended version of Belohlavek, R., and Mikula, T., "Typicality in conceptual structures within the framework of formal concept analysis," Proc. CLA 2020, 33–45. We thank anonymous reviewers for suggesting useful improvements in the conference version of this paper. Supported partly by the project IGA 2020, reg. no. IGA_PrF_2020_019, and by the project IGA 2021, reg. no. IGA_PrF_2021_022, of Palacký University, Olomouc.

Appendix

Instructions of the questionnaire we used to collect the typicality ratings for the Zoo data (the original questionnaire was in Czech):

Hello!
By filling out this questionnaire you contribute to research in the psychology of concepts at the Department of Computer Science, Palacky University Olomouc. The questionnaire takes cca 10 minutes.
You will be asked to assess typicality of animals for three categories (concepts), namely bird, mammal, and fish. Each category shall be assessed on a separate page.
For a given category (e.g. bird), you will see a list of particular animals (birds) of this category. Read the whole list first. Then select for each animal a value in the scale 1 to 5 which describes the extent to which the animal is typical of the category (1 = least typical, 5 = most typical). If you consider it necessary, when filling out the values, go back and change the previously filled values. If you do not know the particular animal, do not select any value (go to the next animal).
When filling out the questionnaire, do not search for additional information (e.g. pictures). Do not spend much time when assigning a typicality value (cca seconds). Do not forget to send out the filled questionnaire. Fill the questionnaire just once.
After sending out your questionnaire, you will be able to see responses of other respondents.
Do not hesitate to contact us if you have questions. Thank you for filling out the questionnaire.

References

[1] L.W. Barsalou, Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories, J. Exp. Psychol. Learn. Mem. Cogn. 11 (4) (1985) 629–654, https://doi.org/10.1037/0278-7393.11.1-4.629.
[2] R. Belohlavek, Fuzzy Relational Systems: Foundations and Principles, Kluwer Academic Publishers, USA, 2002.
[3] R. Belohlavek, J.W. Dauben, G.J. Klir, Fuzzy Logic and Mathematics: A Historical Perspective, Oxford University Press, 2017.
[4] R. Belohlavek, T. Mikula, Zoo typicality, GitHub repository, 2021, https://github.com/mikulatomas/zoo-typicality.
[5] R. Belohlavek, M. Trnecka, Basic level in formal concept analysis: interesting concepts and psychological ramifications, in: Twenty-Third International Joint Conference on Artificial Intelligence, 2013, pp. 1233–1239.
[6] R. Belohlavek, M. Trnecka, Basic level of concepts in formal concept analysis 1: formalization and utilization, Int. J. Gen. Syst. 49 (7) (2020) 689–706.
[7] R. Belohlavek, M. Trnecka, Basic level of concepts and formal concept analysis 2: examination of existing basic level metrics, Int. J. Gen. Syst. 49 (7) (2020) 707–723.
[8] U. Bodenhofer, F. Klawonn, Robust rank correlation coefficients on the basis of fuzzy orderings: initial steps, Mathw. Soft Comput. 15 (2008) 5–20.
[9] C. Carpineto, G. Romano, Concept Data Analysis: Theory and Applications, Wiley, Chichester, England, Hoboken, NJ, 2004.
[10] M.A. Dieciuc, J.R. Folstein, Typicality: stable structures and flexible functions, Psychon. Bull. Rev. 26 (2019) 491–505, https://doi.org/10.3758/s13423-018-1546-2.
[11] S. De Deyne, S. Verheyen, E. Ameel, W. Vanpaemel, M.J. Dry, W. Voorspoels, G. Storms, Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts, Behav. Res. Methods 40 (4) (2008) 1030–1048, https://doi.org/10.3758/BRM.40.4.1030.
[12] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2019, http://archive.ics.uci.edu/ml.
[13] B.S. Everitt, S. Landau, M. Leese, D. Stahl, Cluster Analysis, 5th ed., Wiley, Chichester, West Sussex, UK, 2011.
[14] D.H. Fisher, A computational account of basic level and typicality effects, in: Proc. AAAI-88, 1988, pp. 233–238.
[15] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, Berlin, 1999.
[16] P. Jaccard, Nouvelles recherches sur la distribution florale, Bull. Soc. Vaud. Sci. Nat. 44 (1908) 223–370.
[17] M.-J. Lesot, R. Kruse, Data summarisation by typicality-based clustering for vectorial and non vectorial data, in: FUZZ-IEEE 2006, 2006, pp. 547–554.
[18] E. Machery, 100 years of psychology of concepts: the theoretical notion of concept and its operationalization, Stud. Hist. Philos. Biol. Biomed. Sci. 38 (1) (2007) 63–84, https://doi.org/10.1016/j.shpsc.2006.12.005.
[19] C.B. Mervis, J. Catlin, E. Rosch, Relationships among goodness-of-example, category norms, and word frequency, Bull. Psychon. Soc. 7 (3) (1976) 283–284, https://doi.org/10.3758/BF03337190.
[20] G.L. Murphy, The Big Book of Concepts, MIT Press, Cambridge, Mass., 2002.
[21] H.-S. Park, C.-H. Jun, A simple and fast algorithm for K-medoids clustering, Expert Syst. Appl. 36 (2) (2009) 3336–3341.
[22] E. Rosch, Principles of categorization, in: E. Rosch, B.B. Lloyd (Eds.), Cognition and Categorization, Erlbaum, Hillsdale, NJ, 1978, pp. 27–48.
[23] E. Rosch, C.B. Mervis, Family resemblances: studies in the internal structure of categories, Cogn. Psychol. 7 (4) (1975) 573–605, https://doi.org/10.1016/0010-0285(75)90024-9.
[24] E. Rosch, et al., Basic objects in natural categories, Cogn. Psychol. 8 (3) (1976) 382–439, https://doi.org/10.1016/0010-0285(76)90013-X.
[25] J.P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.
[26] SciPy 1.0 Contributors, et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods (2020), https://doi.org/10.1038/s41592-019-0686-2.
[27] W. Yeh, L.W. Barsalou, The situated nature of concepts, Am. J. Psychol. 119 (3) (2006) 349–384, https://doi.org/10.2307/20445349.
[28] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (3) (1965) 338–353, https://doi.org/10.1016/S0019-9958(65)90241-X.
[29] J. Zhang, Selecting typical instances in instance-based learning, in: Machine Learning Proceedings, Morgan Kaufman, 1992, pp. 470–479.

Appendix B

Similarity metrics vs human judgment of similarity for binary data: Which is best to predict typicality?

Paper (Belohlavek and Mikula, 2024d) describes the evaluation of 69 similarity measures and their ability to predict the typicality ratings. These similarity measures were also compared to the human judgments of similarity provided in the Dutch data. The results of these experiments were briefly described in Section 4.3. The paper resulted from joint research with my supervisor Radim Bělohlávek and was published in Applied Soft Computing.

Applied Soft Computing Journal 153 (2024) 111270

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Similarity metrics vs human judgment of similarity for binary data: Which is best to predict typicality?

Radim Belohlavek, Tomas Mikula *

Department of Computer Science, Palacký University Olomouc, Czech Republic

ARTICLE INFO

Keywords:
Similarity
Similarity measure
Typicality
Binary data

ABSTRACT

Similarity measures for binary data have been subject to a number of comparative studies. In contrast to these studies, we provide a comparison of similarity measures with human judgment of similarity. For this purpose, we utilize the phenomenon of typicality, whose definition is based on similarity. We observe how well the similarity of objects – either computed by a similarity measure or provided by human judgment – enables the prediction of typicality of these objects in various human categories. In doing so, we examine a large variety of existing similarity measures, and utilize recently available extensive data involving binary data as well as data on human judgment of similarity and typicality.

1. Problem description

Measuring similarity of binary data plays a crucial role in many tasks and has been subject to extensive research. Since the first formulas to measure similarity appeared more than a hundred years ago, a multitude of similarity measures, as well as the dual dissimilarity measures, have been proposed in various areas. Given the number of existing similarity measures for binary data, exploration of the proposed similarity measures has naturally become the subject of a number of studies.

The existing works study various properties of the proposed similarity measures, mutual relationships of the measures, and examine the performance of particular similarity measures; see, e.g., [1–7] for some influential as well as recent studies.[1] The comparative studies usually involve tens of similarity measures (see Section 2.1 for details) and the comparison is typically based on evaluating these measures on data from a particular domain of interest, such as biology or chemistry, and also on randomly generated data.

The primary purpose of our paper is different, namely to compare the large variety of the existing similarity measures using extensive psychological data. Such exploration has not been done before and constitutes the main novelty of our contribution. In particular, we compare the available similarity measures on the one hand with a human judgment of similarity on the other hand. We explore this question indirectly via the important phenomenon of typicality, which – according to a common psychological view – is based on the concept of similarity [8]. For this purpose, we utilize our recent results on typicality and its prediction [9]. In particular, we consider the capability of pairwise similarity ratings – those computed by similarity measures and those provided by human judgment – to predict typicality. In addition to relating similarity measures with a human judgment of similarity, our comparison also provides a view on the relationship between the involved similarity measures themselves. Our study is possible due to the now available high-quality psychological data regarding human categories and related phenomena [10], which involves binary data and data on human judgment of similarity and typicality.

In Section 2.1, we present preliminaries on similarity measures; a list of formulas for all the similarity measures involved in our study along with additional information is supplied in the appendix. The phenomenon of typicality and the formula for computing typicality are the subjects of Section 2.2. In Section 3, we describe the data we use in the present study. Our experimental evaluation is the content of Section 4. Section 5 concludes the paper with observations drawn from the experiments.

2. Similarity and typicality

2.1. Similarity measures

For the purpose of our paper, we follow a general understanding according to which a similarity measure on a set X of objects is a binary

* Corresponding author.
E-mail addresses: radim.belohlavek@acm.org (R. Belohlavek), mail@tomasmikula.cz (T. Mikula).

[1] As works on similarity measures for binary data are rather numerous, we only include selected papers, directly related to our purpose, and refer to these papers for further references.

https://doi.org/10.1016/j.asoc.2024.111270
Received 4 August 2023; Received in revised form 5 January 2024; Accepted 13 January 2024
Available online 17 January 2024
1568-4946/

function

$sim : X \times X \to \mathbb{R};$

the value $sim(x, y)$ is interpreted as the extent to which x is similar to y. This general approach subsumes a variety of particular similarity measures proposed in the literature. That is, we do not impose possible additional constraints, such as $sim(x, y) = sim(y, x)$, $sim(x, y) \leq sim(x, x)$, or various dual forms of the triangle inequality, which are sometimes considered.

When the similarity of binary data is considered, the set X consists of all possible objects described by n binary attributes, and may hence be conveniently identified with the set $\{0, 1\}^n$ of all n-dimensional binary vectors. Thus, for instance,

$x = \langle 1, 0, 0, 1, 1 \rangle$

represents an object described by 5 binary attributes, i.e., $x \in \{0, 1\}^5$, and one has $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, $x_4 = 1$, and $x_5 = 1$. That is, the object has the first, the fourth, and the fifth attribute, but not the second, nor the third.

The similarity measures considered in the literature can conveniently be defined in terms of the values a, b, c, and d, defined as follows. Consider n attributes and two binary vectors $x, y \in \{0, 1\}^n$, and let

$$
\begin{aligned}
a &= \#\{i \mid x_i = 1 \text{ and } y_i = 1\},\\
b &= \#\{i \mid x_i = 1 \text{ and } y_i = 0\},\\
c &= \#\{i \mid x_i = 0 \text{ and } y_i = 1\},\\
d &= \#\{i \mid x_i = 0 \text{ and } y_i = 0\}.
\end{aligned}
$$

That is, a is the number of common presences and d is the number of common absences of the attributes $i = 1, \ldots, n$. On the other hand, b is the number of attributes present on x but not on y, and c is the number of attributes absent on x but present on y. While a and d indicate similarity of x and y, b and d indicate dissimilarity. Clearly, $a + b + c + d = n$.

For example, for the vectors

$$
\begin{aligned}
x &= \langle 0, 1, 1, 0, 0, 1, 0, 1, 1, 0 \rangle,\\
y &= \langle 1, 1, 1, 0, 0, 1, 0, 1, 0, 1 \rangle
\end{aligned}
$$

in $\{0, 1\}^{10}$, one has

$$
\begin{matrix} a = 4 & b = 1 \\ c = 2 & d = 3 \end{matrix} .
$$

Now, a similarity measure may be defined by a formula involving the coefficients a, b, c, and d, corresponding to $x, y \in \{0, 1\}^n$, such as

$$
sim(x, y) = \frac{a + d}{a + b + c + d} \qquad \text{and} \qquad sim(x, y) = \frac{a}{a + b + c}. \tag{1}
$$

The formulas in (1) actually represent two well-known similarity measures, the simple matching coefficient (SMC) and the Jaccard measure (Jac), respectively.

Measures of similarity for binary data have a long history; see, e.g., [1,2,4]. The first measures were proposed at the end of the 19th century to facilitate the analysis of biological species, which were often described in terms of binary attributes. Since then, numerous other measures have been proposed in areas as diverse as biology, ecology, geology, psychology, chemistry, medicine, information retrieval, machine learning, and bioinformatics. A principal reason for the continuing interest in these measures is the omnipresence of data describing various kinds of items, such as biological species, chemical compounds, performance tests, or documents, in terms of binary attributes, and the need to analyze such data.

Even though no definite categorization or grouping of similarity measures for binary data has been established in the literature, a few classification criteria have been considered. The following two seem best known. The first one attempts to classify the measures into statistically based and co-occurrence based. The statistically based, also called association measures, are often interpretable as correlation coefficients and have usually their values in the interval $[-1, 1]$. Their formulas may seem less intuitive and often contain $ad - bc$ in the numerator. The co-occurrence-based measures are based on the frequencies a and d of co-occurrence of the involved binary attributes, have their values in $[0, 1]$, and are usually defined by intuitive formulas such as (1), which contain a or $a + d$ in the numerator. The second widely used criterion consists in whether the measure takes into account, ignores, or takes into account partially the number d of common absences (negative matches) of the attributes. For instance, while the above SMC measure takes d into account in that it increases similarity, the Jaccard measure ignores d.

In our study, we employ 69 similarity measures, which we selected from a large variety of similarity measures described in the literature, particularly in [1,2,4–6]. The employed measures, along with comments on the logic of our selection and further information about these measures are described in the appendix of this paper. In particular, a list of the employed measures is provided by Table 4, which contains an abbreviation and a name for each measure as well as a formula for computing the values of a given measure. The list is sorted alphabetically by the abbreviations so that a reader may quickly find details about the measures when assessing our experimental results.

2.2. Typicality

The phenomenon of typicality is well known from everyday life: Intuitively, a sparrow is a typical bird, an ostrich is not. Typicality is one of the most important phenomena accompanying human concepts and plays a significant role in a variety of cognitive tasks including categorization and classification. Since being typical is a matter of degree, typicality manifests a graded structure of concepts. Both typicality and the graded structure of concepts have been among the central topics of research in the psychology of concepts since the 1970s. For a comprehensive exposition of typicality and its role in the psychology of concepts, we refer to [8].

According to a mainstream psychological view, which goes back to the seminal work by Eleanor Rosch and her colleagues [11–13], the notion of typicality of an object in a concept (category) is based on the notion of similarity: An object is considered typical in a given concept if the object is similar to the objects to which the concept applies. In [9], we formalized this view of Rosch as follows[2]:

Definition 1. Given a similarity $sim : X \times X \to \mathbb{R}$, an object $x \in X$, and a nonempty set $A \subseteq X$ representing a concept, a *degree of typicality of x in A* is defined by

$$
typ(x, A) = \frac{\sum_{x_1 \in A} sim(x, x_1)}{|A|}. \tag{2}
$$

Formula (2) for typicality results as a straightforward formalization of a verbal description of the psychological view available in the literature and represents the average similarity of the object x to all the objects in A. As demonstrated in [9], the degrees of typicality computed by this formula are highly correlated with human judgment of typicality, i.e., with degrees of typicality provided by humans.

Example 1. Table 1 presents a part of the Zoo data [14], restricted to 9 exemplars of the category "bird" (sparrow, …, penguin) and some of their binary attributes (feathers, …, legs 2). The column labeled

[2] In fact, in [9] we used formula (2) to define typicality of x in A, for A being an extent of a so-called formal concept. The definition in the present paper simply gets rid of the constraint and allows A to be a general subset of X, i.e., allows A to represent an arbitrary category.

Table 1
Values of typicality of exemplars of the category "bird" from example 1.

	feathers	eggs	airborne	aquatic	predator	backbone	breathes	tail	domestic	catsize	legs 2	typ(J)
sparrow	1	1	1	0	0	1	1	1	0	0	1	0.809
crow	1	1	1	0	1	1	1	1	0	0	1	0.807
vulture	1	1	1	0	1	1	1	1	0	1	1	0.802
duck	1	1	1	1	0	1	1	1	0	0	1	0.784
swan	1	1	1	1	0	1	1	1	0	1	1	0.783
kiwi	1	1	0	0	1	1	1	1	0	0	1	0.763
ostrich	1	1	0	0	0	1	1	1	0	1	1	0.759
chicken	1	1	1	0	0	1	1	1	1	0	1	0.745
penguin	1	1	0	1	1	1	1	1	0	1	1	0.745

typ(J) provides the values of typicality of the exemplars, i.e., the values $typ(x, A)$ computed according to (2), based on the Jaccard similarity (cf. Section 2.1). Note that in (2), x denotes the exemplar whose typicality is being computed, A represents the 9-element set of exemplars of "bird," and $sim(x, x_1)$ denotes the Jaccard similarity of exemplars x and x_1 calculated from the binary descriptions of the two exemplars provided by the corresponding table rows.

The ordering of the exemplars in the table by the values of typicality corresponds to intuition despite the limited number of attributes used in our illustrative example; see [9] for a more comprehensive study of typicality in the context of the Zoo data. Note also that the relatively low dispersion of typicality values results from the limited number of the exemplars and attributes involved in this illustrative example.

3. Data

The availability of high-quality data is essential for any kind of experiment that aims to be psychologically relevant. For our purpose, the Dutch data [10] is unique in this regard, as it provides perhaps the most comprehensive data regarding common human categories and their numerous characteristics, including similarity and typicality. Moreover, the data is considerably larger than the previously available psychological data of similar nature. In this section, we provide a brief description of the data, particularly the parts we use, and our comments regarding usability in experiments along with our technical modifications in this regard.

The Dutch data has been gathered by psychologists at the University of Leuven in a thorough, carefully designed study involving hundreds of human respondents. It basically provides information regarding common language concepts (categories), binary attributes (features) relevant to these categories, objects (exemplars) in these categories, and various psychologically relevant characteristics.

In particular, the data involves 16 linguistic categories. These include both the so-called natural kind and artifact categories, as these two kinds are commonly believed to have distinct properties. Each category is represented by a number of objects (exemplars), such as a robin for the category "bird." There are 10 natural kind categories: "fruit" (30 exemplars); "vegetables" (30); "professions" (30); "sports" (30); the animal categories "amphibians"(5),[3] "birds" (30), "fish" (23), "insects" (26), "mammals" (30), and "reptiles" (22).[4] In addition, there are 6 artifact categories: "clothing" (29), "kitchen utensils" (33), "musical instruments" (27), "tools" (30), "vehicles" (30), and "weapons" (20).[5]

These categories comprise 249 exemplars for the natural kind and 166 exemplars for the artifact categories, which were obtained from humans and are representative of the respective categories.[6] Coverage by these categories is considerable; for instance, the animal categories cover a rather large part of the known animal domain. The objects (exemplars) and attributes (features) were obtained by processes described in [10]. In particular, the attributes were generated by 1003 respondents in two ways: First, respondents were asked to list relevant attributes for a given category (these are called category attributes). Second, they were asked to list relevant attributes for each object involved in the data (these are called exemplar attributes). Furthermore, unions of all exemplar features listed for all the objects in a given category were considered, as well as the union of all exemplar features of all the objects in the animal domain, and an analogous union of exemplar features for the artifact domain.

An essential part of the data are the so-called exemplar-by-feature applicability matrices. These are various matrices in which the rows and columns correspond to some of the objects and attributes, respectively, and the entries contain information about whether a particular object has or does not have a particular attribute. Each of the matrices was filled separately by four respondents. The data also contains the corresponding aggregated matrices, in which the values, viz. 0, 1, 2, 3, and 4, indicate the number of respondents who agreed on that the respective object has the respective attribute. To obtain binary matrices (and thus data with binary attributes) from these aggregated matrices, one naturally thresholds the matrix entries. We present our experiments for a threshold equal to 2. Hence, our binary matrices contain 1 in the entry corresponding to the object x and the attribute y if at least two respondents agreed that x has y.

In particular, we use the binary matrices described in Tables 2 and 3. For instance, the first row in Table 2 refers to two binary matrices: The first one, a 30×28 matrix, describes which of the 30 exemplars of the category "bird" has which of the 28 category attributes of this category (i.e., attributes listed as category attributes for this category by respondents); the second one, a 30×225 matrix, describes which of the 30 exemplars of the category "bird" has which of the 225 exemplar attributes for this category (i.e., all attributes listed as exemplar attributes for some exemplar of "bird"). Similarly, the 129×225 binary matrix referred to by the first row in Table 3 describes which of the 129 objects in the animal domain have which of the corresponding 225 category attributes; the 129 objects are all the objects of the categories "amphibian", "bird", "fish", "insect", "mammal", and "reptile", and the 225 category attributes are all attributes listed as category attributes for these six categories. Likewise, the 129×764 matrix describes which of the objects in the animal domain have which of the corresponding 764 exemplar attributes, i.e., all the attributes listed as exemplar attributes for some of the 129 exemplars in the animal domain.

Typicality ratings, which are present in the Dutch data, were obtained from 112 respondents. For each of the 16 categories and each object in the respective category, the data contains a typicality rating on the scale 1 (very atypical) to 20 (very typical).

The pairwise similarity ratings of the Dutch data come partly from the previous study [15], in which the ratings were obtained for ten of the present categories from 42 participants. The ratings for the other categories were obtained from 92 respondents in [10], who also

[3] Since the category "amphibians" only contains 5 exemplars, and since these exemplars are included in the category "reptiles," we omit it in most of our considerations below; see [10] for reasons to include the exemplars of "amphibians" in "reptiles."

[4] The exemplar-by-feature applicability matrices, which we describe below and use in our experiments, contain only 20 exemplars of the category "reptiles," because the respondents who were to fill in these matrices turned out to not to be familiar with two exemplars (komodo and iguanodon). We hence exclude these two exemplars from our experiments.

[5] Here, we use plural in category names, as the authors do [10]; below, we use singular, i.e., "bird" rather than "birds" to be consistent with our previous writings.

[6] In addition to the 5 amphibians included in reptiles and two omitted exemplars of reptiles (see above), note that three exemplars of artifact categories are included in two distinct categories.

Table 2
Category-based binary matrices used in our experiments.

Category	Objects	Category attributes	Exemplar attributes
bird	30	28	225
clothing	29	38	258
fruit	30	32	233
fish	23	32	156
insect	26	37	214
kitchen utensil	33	39	328
mammal	30	34	288
musical instrument	27	39	218
profession	30	21	370
reptile	20	35	179
sport	30	33	382
tool	30	37	285
vegetable	30	30	291
vehicle	30	34	322
weapon	20	32	181

Table 3
Domain-based binary matrices used in our experiments.

Domain	Objects	Category attributes	Exemplar attributes
animal	129	225	764
artifact	166	301	1,295

provided additional ratings for the ten categories involved in [15] to improve reliability. For every category – except for "amphibians," whose five exemplars are included in "reptiles" – and each pair of objects, the data contains a similarity rating on the scale 1 (totally dissimilar) and 20 (totally similar).

Since the original data contains some minor semantic and technical faults, as well as inconveniences as regards a possible machine processing of the data, we modified the data as follows. For one, since the original data contains some wrongly formatted comma-separated files, we transformed them into a valid format. In addition, the names of some objects and attributes are spelled differently across multiple files in the original data; we therefore unified these names. We also converted all names to lowercase to unify them. No changes were made to the data itself. The result is easily machine-processable data. The corrected version of Dutch data, along with a convenient Python wrapper, is publicly available on GitHub [16].

4. Experiments

4.1. Rationale

Comparing similarities via the ability to predict typicality. The rationale of our experiments may be described as follows. Formula (2) for computing degrees of typicality involves degrees $sim(x, y)$ of similarity. Hence, for a given similarity function sim, the function typ may be regarded as a function $typ(sim)$ parameterized by sim, which assigns to each x in a given universe X of objects and a non-empty subset A of X the degree

$$[typ(sim)](x, A) = \frac{\sum_{x_1 \in A} sim(x, x_1)}{|A|}$$

to which the object x is typical for the concept (category) represented by A.

As explained in Section 3, the Dutch data contains information regarding the objects (exemplars) of a variety of categories, including descriptions of these objects by binary attributes. The descriptions of objects by binary attributes enable one to compute the values $sim(x, y)$ of similarity measures sim for pairs of objects x and y. Consequently, one may compute, for any given category A, the degrees $[typ(sim)](x, A)$ of typicality determined by each particular similarity measure sim. In addition, since the Dutch data also contains information on human judgment of similarity, i.e., contains similarity degrees $\mathrm{HJ}(x, y)$ obtained from humans for pairs of the involved objects x and y, one may also compute the degrees $[typ(\mathrm{HJ})](x, A)$ of typicality determined by human judgment of similarity HJ. From this perspective, different similarities shall generally lead to different predictions of typicality.

Now, since the Dutch data also contains degrees of typicality assessed by humans for the involved categories, one may explore, for a given category A and for each similarity measure sim, a correlation of the computed typicality degrees $[typ(sim)](x, A)$ for the objects x in A on the one hand, and the degrees of typicality obtained for the category A from humans on the other hand. The same kind of correlation may be explored for the typicality degrees $[typ(\mathrm{HJ})](x, A)$ computed using human similarity in place of $[typ(sim)](x, A)$. High correlation implies that the particular similarity (represented by a similarity measure or by human judgment) is capable of predicting well the human judgment of typicality.

One may then explore various questions; most importantly:

- How do the various similarity measures compare in their ability to predict typicality?
- How do the similarity measures compare to a human similarity in the same regard, i.e., in their ability to predict typicality?

It is basically these questions that we examine using the experiments presented below. Note that while various comparisons of selected similarity measures are available in the literature (cf. Section 2.1), comparing similarity measures with human judgment of similarity has never been explored in the literature.

Assessment of correlation. The design of our experiments implies a need to assess correlation in the following scenario. For a given category A and a given similarity function sim (either a similarity measure or a similarity obtained from human judgment), we need to assess the correlation between a typicality rating of objects (exemplars) x of the given category, computed by the above formula for $[typ(sim)](x, A)$, and a typicality rating given by a human judgment. To assess correlation of these two typicality ratings, we use the well-known Kendall tau rank-order correlation coefficient.

Recall that the Kendall-tau coefficient measures agreement between two linear orderings (rank orderings), $<_1$ and $<_2$, on a given set of objects. Its basic version is defined by

$$\frac{\#\text{ concordant pairs} - \#\text{ discordant pairs}}{\#\text{ all pairs}};$$

here, a pair of objects x and y is concordant if $x <_1 y$ and $x <_2 y$, or $x >_1 y$ and $x >_2 y$, and is discordant if $x <_1 y$ and $x >_2 y$, or $x >_1 y$ and $x <_2 y$.

In our scenario, the first ordering of the objects, $<_1$, is determined by the computed typicality $typ(sim)$, while the second one, $<_2$, is given by the human rating of typicality, and the Kendall tau is applied to these orderings. In this sense, Kendall tau measures the extent to which the typicality rating determined by the chosen similarity sim agrees with the typicality rating given by human judgment.

Note also that we chose the τ_b variant of the Kendall coefficient since it properly accounts for ties, i.e., situations in which the same degree of typicality is assigned to two or more objects. The coefficient τ_b ranges from 1 (same ordering) to −1 (inverse, i.e., opposite ordering). We used the implementation of τ_b in a Python library [17].

4.2. Results

Our first set of experiments involves the category-based matrices described in Table 2. As described in Section 3, each of these 30 matrices corresponds to a single category and one of the two kinds of attributes (category and exemplar). For each such matrix and each considered similarity sim (i.e., each considered similarity measure of Table 4 and the human similarity obtained from the Dutch data), we computed the degrees $[typ(sim)](x, A)$ of typicality for all objects x of the respective category (i.e., for all matrix rows).[7] We then computed the Kendal τ_b correlation coefficient of the computed degrees of typicality and the human-assessed typicality degrees for the given category. The results for all the natural kind categories and their category attributes are displayed in Fig. 1. Fig. 2 shows analogous results for the exemplar attributes. The results for all the artifact categories and their category and exemplar attributes are shown in Figs. 3 and 4, respectively.

In this and the other graphs, we use the abbreviations introduced in the appendix (Table 4) to denote the respective similarity measures. Thus, for instance, typ(Di2) denotes the typicality computed by means of the Di2 (Dice 2) similarity measure. In the same spirit, typ(HJ) denotes the typicality computed by means of the human judgment of similarity. The $typ(sim)$ on the horizontal axis are ordered by the mean value of the correlation coefficients across the involved categories.

The second set of experiments involves the four domain-based matrices of Table 3. We performed analogous computations as in the first set of experiments. First, for each category in the animal domain, we computed the degrees of typicality using all the category attributes of the domain matrix for each object of the category. Then a Kendall τ_b coefficient of the computed typicality degrees and the human-assessed degrees of typicality was computed for each particular category. The results are displayed in Fig. 5. The results of the same computation with all the exemplar attributes of the animal domain replacing the category attributes are shown in Fig. 6. Analogous results for the artifact domain and its categories are presented in Figs. 7 and 8. Notice that the categories "fruit", "profession", "sport", and "vegetable" are not included in the second set of experiments because these categories are not part of the two domains.

To provide a summarized view of the results, we also include Figs. 9, 10, and 11, which display the average correlation coefficients over all the categories in the animal domain, the artifact domain, and in both of these domains, respectively. In each graph, the mean correlation coefficients are presented for the four sets of attributes: the category-based category attributes, the category-based exemplar attributes, the domain-based category attributes, and the domain-based exemplar attributes; cf. Tables 2 and 3.

4.3. Discussion

Both the detailed graphs (Figs. 1–8) and the averaged summary views (Figs. 9–11) reveal notable patterns as regards the ability to predict human judgment of typicality by various similarity functions, as well as regards a comparison of the explored similarity measures and human judgments of similarity. Note first that according to a commonly accepted interpretation, the values of τ_b of rank-order correlation may be interpreted as follows: $\tau_b \geq 0.3$, $0.2 \leq \tau_b < 0.3$, $0.1 \leq \tau_b < 0.2$, and $0.0 \leq \tau_b < 0.1$ indicate strong, moderate, weak, and very weak correlation, respectively; the negative values of τ_b are interpreted analogously.

Consider first the human similarity HJ. Overall, HJ enables rather good predictions of typicality and is among the best similarities in this regard. Not only ranks the human similarity as the sixth best as regards average of correlations across all the categories and all the sets of attributes (Fig. 11) with a rather strong $\tau_b = 0.42$, but performs best as regards prediction of typicality in the animal domain (Fig. 9).

The slightly worse performance of human similarity on the artifact categories and also on the three natural categories outside the animal domain may, in our view, be due to the fact that a judgment of similarity of exemplars of these categories is somewhat problematic (consider, e.g.: What is the similarity degree of sailing and sport fishing, of being an accountant and a postman, sled and bicycle?) and the calculated similarity may hence yield better predictions of typicality.[8]

As regards the performance of all the involved similarities, the averaged summary graph (Fig. 11) indicates that there is a group of similarities with an overall strong correlation of human judgment of typicality. Naturally, this group does not have a sharp boundary, but among its core members are, except for the human similarity HJ discussed above, the similarity measures Co1, RR, int, Di2, and CT3, which all have higher average correlation compared to HJ across all categories and across the artifact domain (Fig. 10). In addition, there is a group of other highly correlated similarity measures, which include Fai, FM, CT4, Fos, Ku2, McC, Sor, SS1, cos, Jac, Maa, and Gle.

Observe that some of the similarity measures display a high average correlation except for predictions in the category-based data with category attributes. We contend that the latter drop in correlation is mainly due to the fact that the category attributes of the smaller, category-based matrices provide less information about the exemplars—a significant phenomenon to which we turn below.

One can also identify a group of similarity measures with a low average correlation and with values around 0, and varying considerably in prediction of typicality over the domain-based and the artifact-based data and the two respective kinds of attributes. These include Den, Co2, Col, Di1, Twd, Fo1, and Gow. From this point of view, Gow seems particularly peculiar as its correlation attains significant negative values in several cases but not in others, which is apparent in all figures except Figs. 1 and 3.

Worth noting is also the good prediction of typicality by Co1 and Di2, and the poor performance of their symmetric counterparts, Co2 and Di1. In both cases, good prediction results when the value of c (see Section 2.1) increases the value of the denominator in the respective similarity formula; hence, if the value $sim(x, x_1)$ involved in formula (2) for typicality gets smaller when x does not have an attribute possessed by x_1 but does not get smaller when x has an attribute not possessed by x_1.

Note at this point that as may be observed in the graphs, certain groups of similarity measures displayed a perfect correlation τ_b in that the correlation coefficient with a human judgment of typicality is the same for all data we explored. This pertains to the pairs BU1 and BU2, Gle and Maa, Ku2 and McC, RG and Sco, and to the triplet Ham, ip, and SMC. In all these cases the respective pairs of similarity measures yield different values, i.e., are distinct functions. Their formulas are, nevertheless, closely related.

Another conclusion which may be drawn from the experiments pertains the quality of attributes. It is well known in the psychology

[7] The similarity measures with undefined values are not included; see Remark 1 in the appendix.

[8] See Section 5 for more details. Human similarity HJ was assessed by the respondents with no context, in that each respondent was asked to judge the similarity for a number of exemplar pairs selected across various categories. We hypothesize that such assessment yields different, likely smaller and less consistent, degrees of similarity compared to an alternative scenario, in which a category name and a list of all exemplars of the category are given, and the respondent is to assess similarity of all exemplar pairs in this category. The name and the list of all objects of the category provide a context for the assessment. In the presence of this context, the assessed similarity degree of, e.g., sled and bicycle, is likely to be higher compared to when no context is present (the context helps one realize, so to say, the similarity because relevant attributes become more apparent in the presence of the context). When assessing typicality, respondents implicitly utilize their context-based judgment of similarity (because then, the category name and the lists of exemplars are available). Now, we hypothesize that the similarity computed using a reasonably good similarity measure sim is likely to be better correlated with the context-based human similarity rather than with the without-context similarity HJ. Hence, the correlation of the human typicality rating with typ(sim) is likely to be higher than the correlation with typ(HJ). This hypothesis would hence explain the slightly worse correlation of the computed typicality based on human similarity compared to computed typicality based on a reasonably good similarity measure.

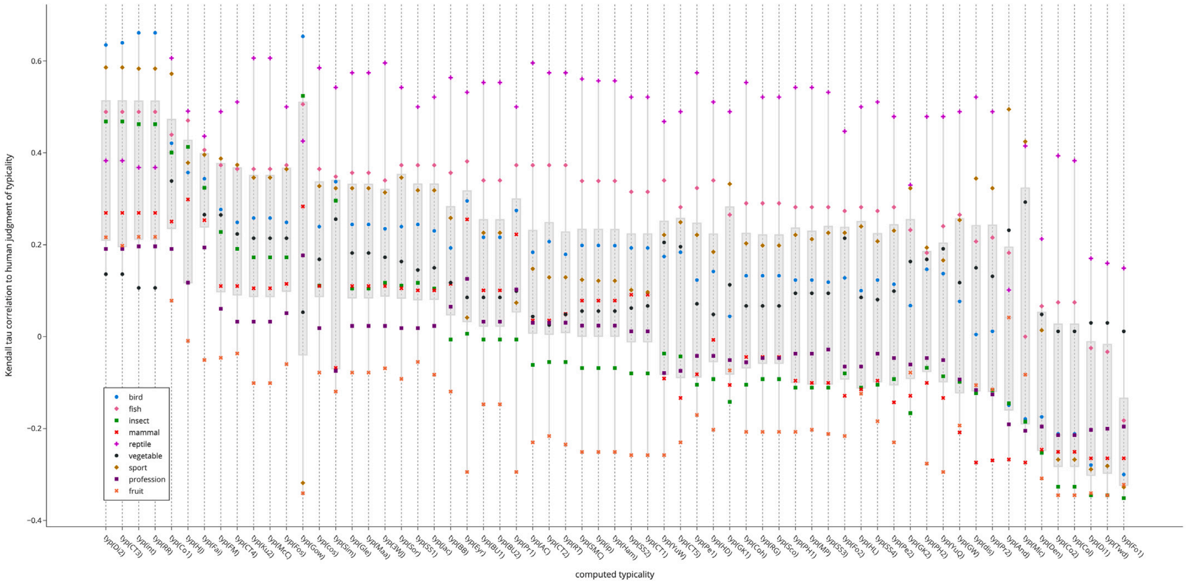

Fig. 1. Correlations of computed typicality to human judgment of typicality across natural categories with category attributes (horizontal axis ordered by mean value).

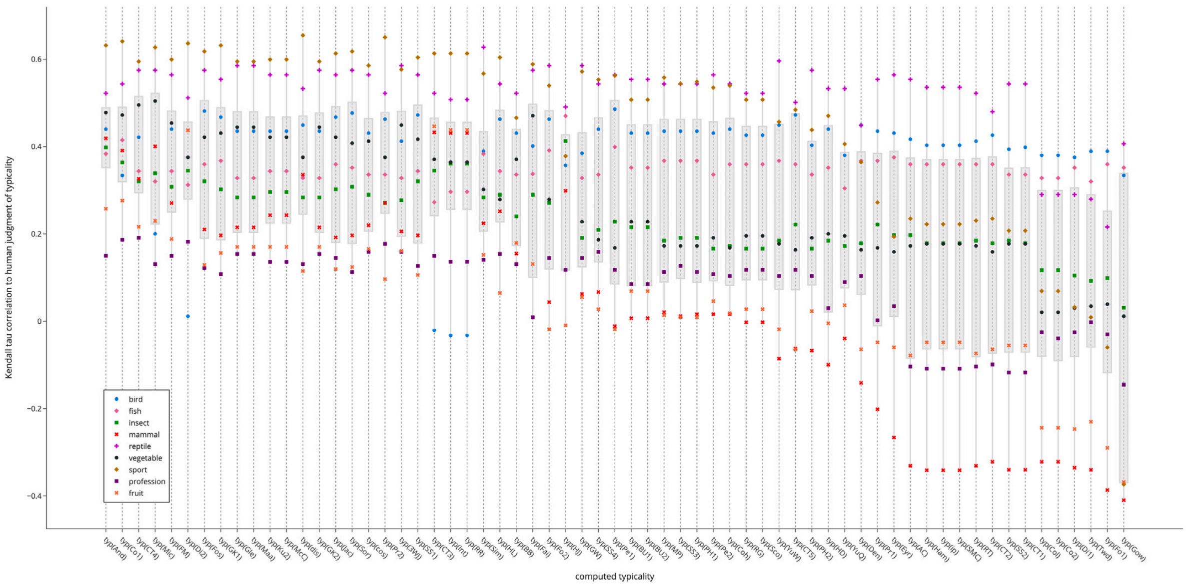

Fig. 2. Correlations of computed typicality to human judgment of typicality across natural categories with exemplar attributes (horizontal axis ordered by mean value).

of concepts that the quality of attributes used to assess typicality and similarity is essential [8]; see also [10] and the references therein. In order to enable good predictions, the attributes need to represent well the aspects people naturally take into account in their judgments on typicality and similarity. This intuitive knowledge has, nevertheless, not been confirmed by any extensive experimentation. Our results provide confirmation of this knowledge. Namely, as is apparent from all the graphs, the exemplar attributes generally result in a better prediction of human judgment of typicality than the category attributes, which are considerably less numerous and provide less distinctive information about the exemplars due to how these kinds of attributes have been collected (see Section 3). This is particularly apparent for the category-based data with the category attributes because, for this data, the numbers of attributes are considerably smaller than for the corresponding data with the exemplar attributes and also much smaller than the numbers of the exemplar and category attributes for the domain-based data. For the domain-based data, the numbers of both kinds of attributes are rather high, resulting in a comparable performance of prediction in this case.

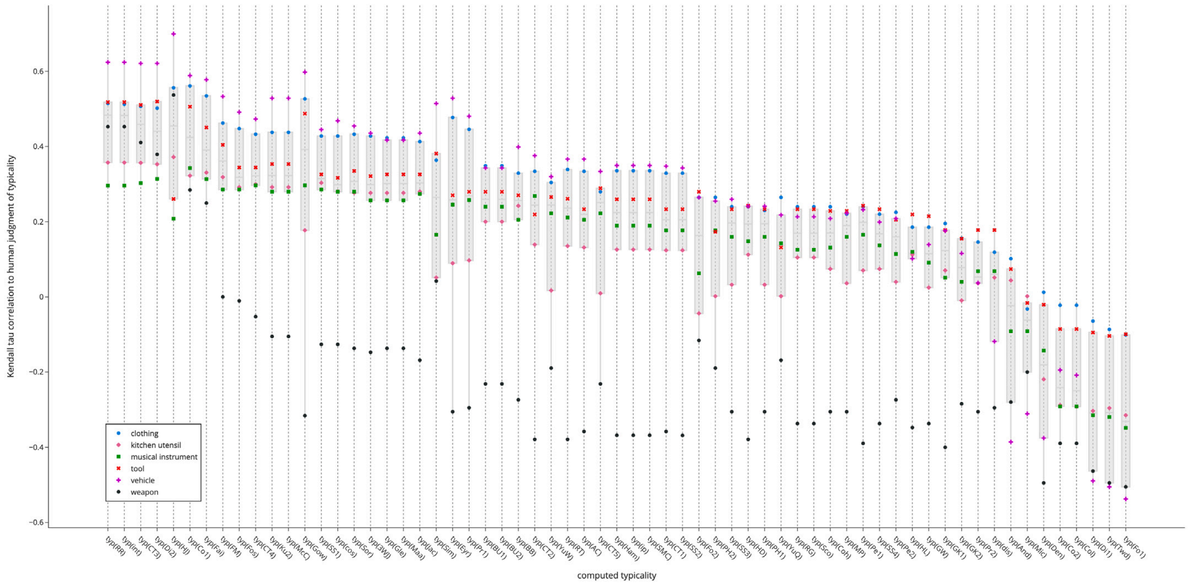

Fig. 3. Correlations of computed typicality to human judgment of typicality across artifact categories with category attributes (horizontal axis ordered by mean value).

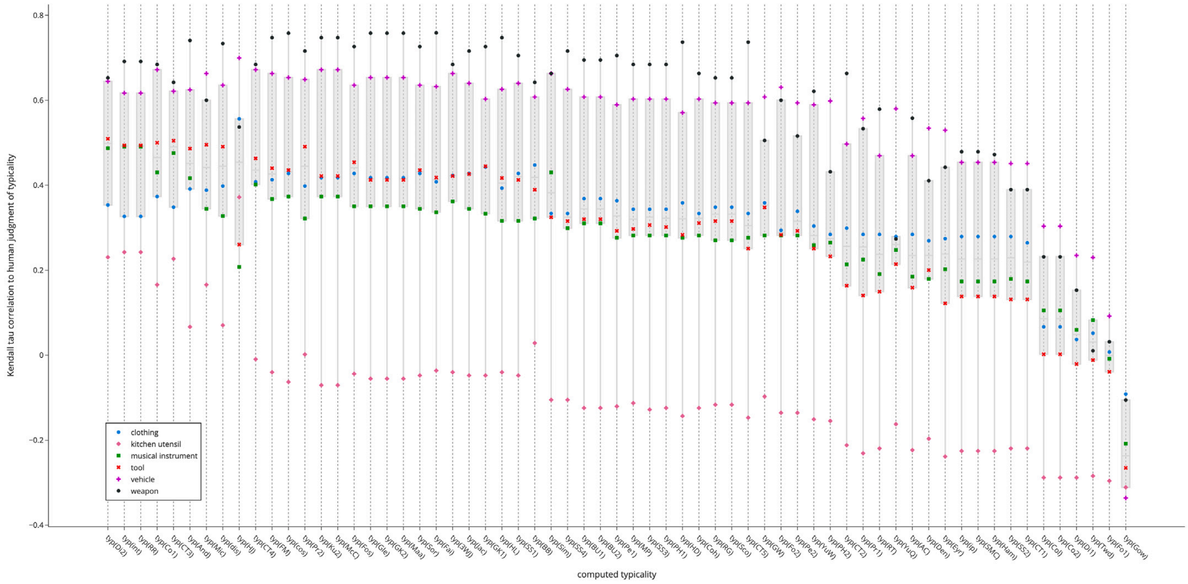

Fig. 4. Correlations of computed typicality to human judgment of typicality across artifact categories with exemplar attributes (horizontal axis ordered by mean value).

As regards a possible answer to the question in the title of our paper, i.e., which similarity is best to predict typicality, it comes as no surprise that there is no clear winner. This seems to result from the fact that all the similarity measures have been carefully designed to serve in certain real situations and have been proven through the test of time. In addition, several measures have been proposed for each particular purpose in the past. It is hence to be expected that groups of similarities, albeit vaguely delineated, rather than a single similarity, might be identified as the best predictors of typicality. In this regard, the group consisting of Co1, RR, int, Di2, CT3, and HJ may be identified as representing the best predictors. It is significant that this group includes the human similarity HJ, which not only confirms an intuitive expectation (human similarity is expected to come out among the best similarities) but also justifies the adequacy of formula (2) for computing degrees of typicality (the formula provides a verified relationship between a human judgment of similarity and a human judgment of typicality). As regards possible common properties of Co1, RR, int, Di2, and CT3, except for Co1, they are examples of the co-occurrence similarity measures defined by intuitive formulas. Moreover, the number d of negative matches (see Section 2.1) does not increase the value of similarity for these measures. We do not have an intuitive explanation for the good performance of the statistically

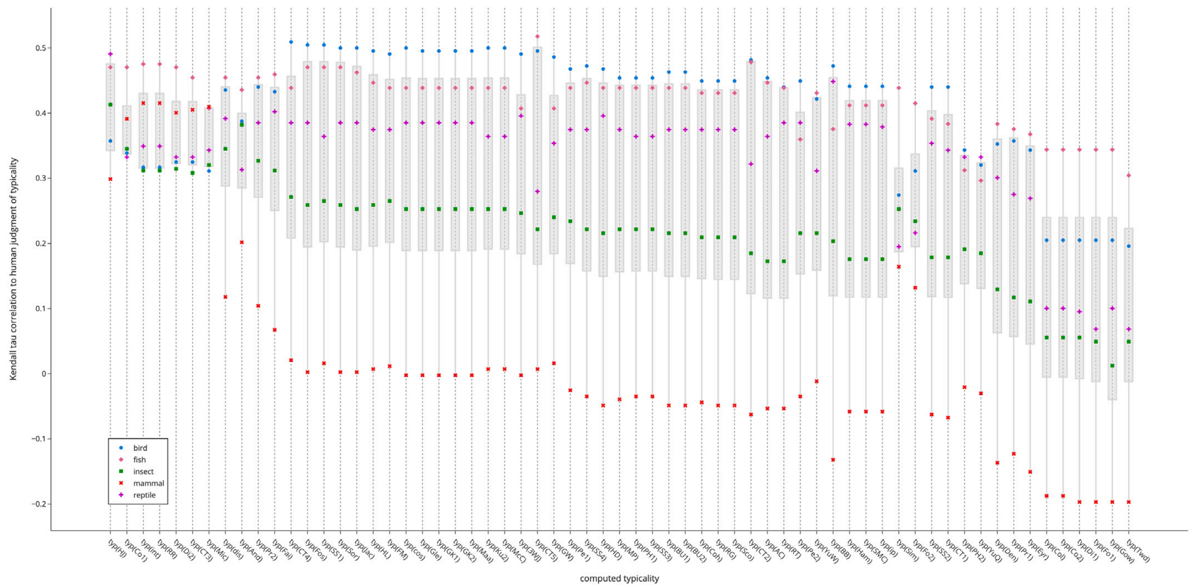

Fig. 5. Correlations of computed typicality to human judgment of typicality across animal domain with category attributes (horizontal axis ordered by mean value).

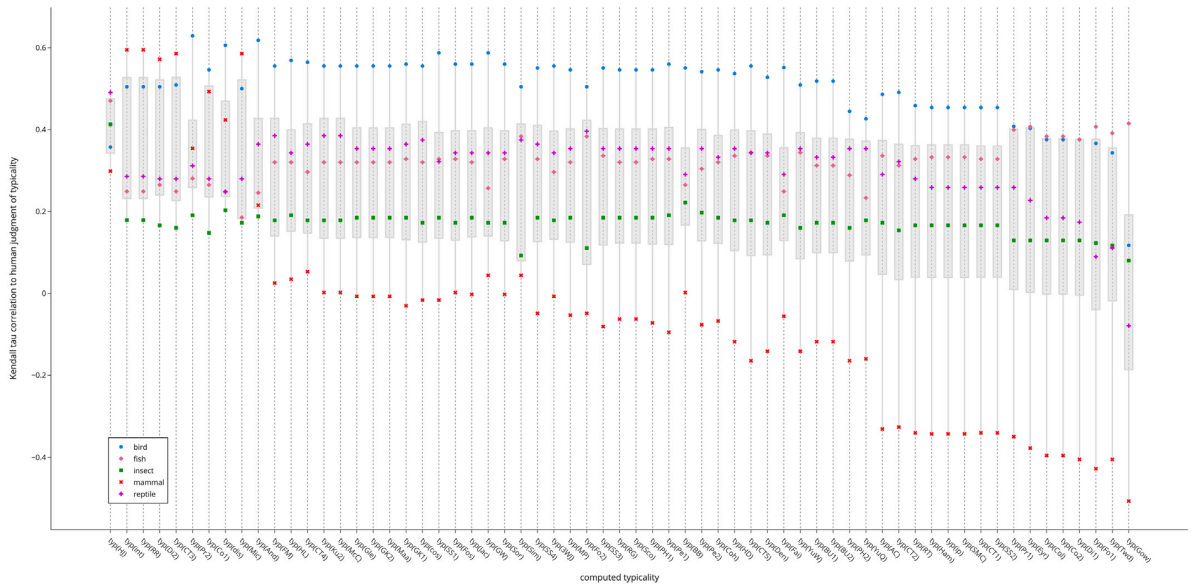

Fig. 6. Correlations of computed typicality to human judgment of typicality across animal domain with exemplar attributes (horizontal axis ordered by mean value).

motivated Co1. The second group that still provides very good predictions of typicality consists of Fai, FM, CT4, Fos, Ku2, McC, Sor, SS1, cos, Jac, Maa, and Gle. A majority of these measures are also co-occurrence based and for all of them, except for Fai, the negative matches (d) do not increase the value of similarity. On the other hand, similarities Den, Gow, Co2, Col, Di1, Twd, and Fo1 lead to poor predictions of typicality. Except for Di1, these are statistically motivated measures and for most of them, the negative matches (d) do increase the similarity value.

The graphs also reveal a few interesting particular observations. For instance, Figs. 3 and 4 display that for the category "weapon," the exemplar attributes result in the best predictions of typicality across all the artifact categories (with correlation values around 0.7), while the category attributes for "weapon" result in the worst prediction, and this holds true for most of the similarity measures. This is likely to be attributed to the small number of category attributes for this category, which turn out poorly informative for the prediction of typicality with most of the measures. We refrain from a detailed exposition of such particular observations, however interesting they may be and leave them for possible future examination due to lack of space.

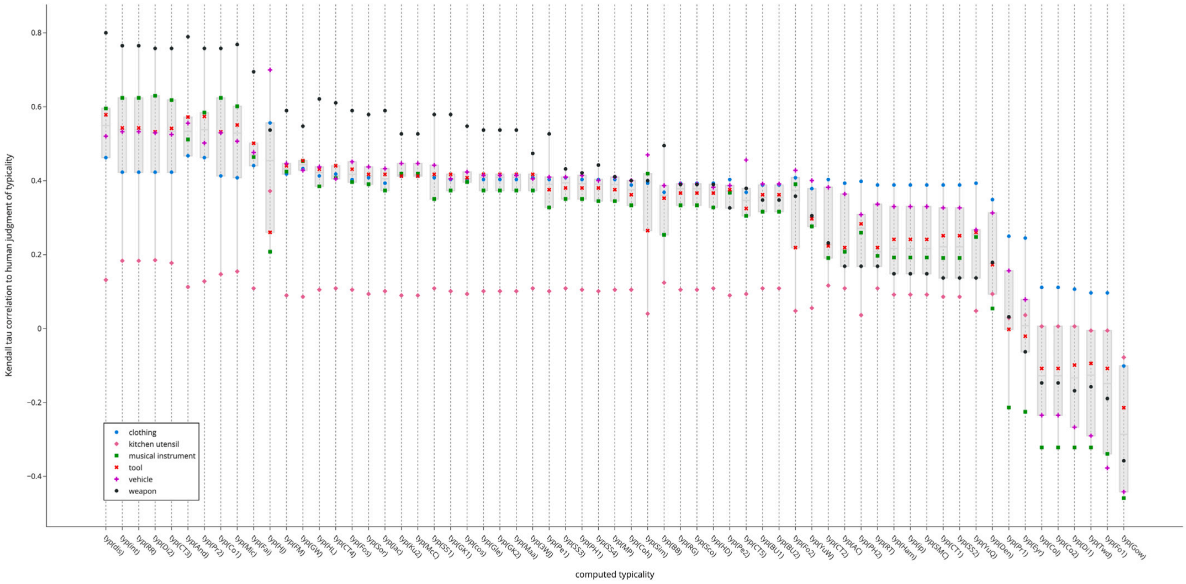

Fig. 7. Correlations of computed typicality to human judgment of typicality across artifact domain with category attributes (horizontal axis ordered by mean value).

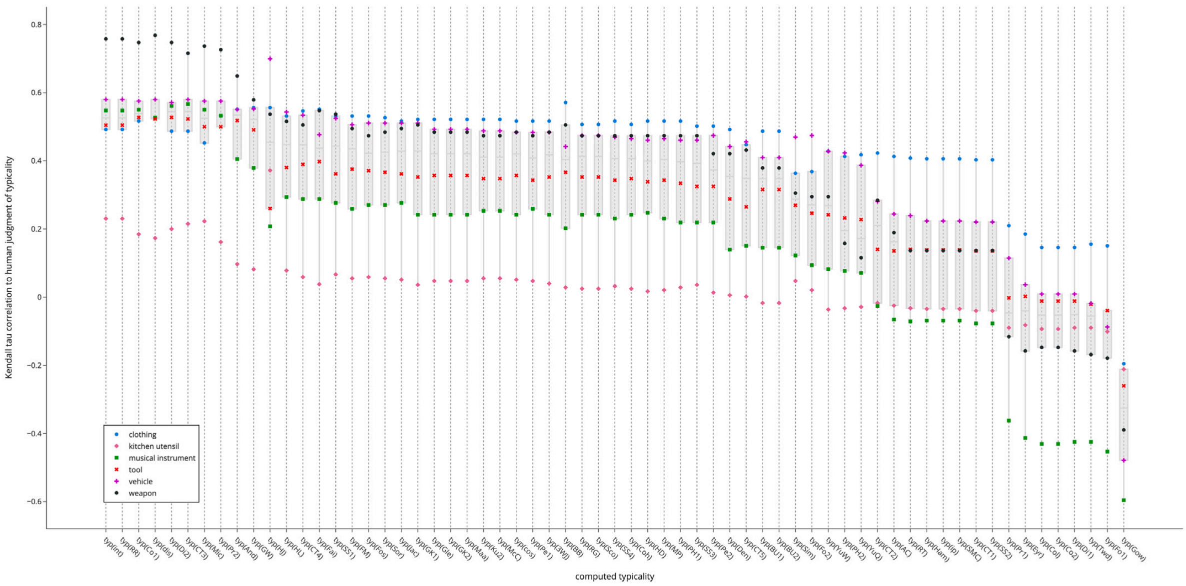

Fig. 8. Correlations of computed typicality to human judgment of typicality across artifact domain with exemplar attributes (horizontal axis ordered by mean value).

As regards possible limitations of the conclusions drawn from our experiments, they are implied, for the most part in our view, by the nature of the test data we use. For one, even though the Dutch data we utilized is rather extensive and involves several binary matrices, which we used, the validity of our conclusions would be improved if supported on yet another data, i.e., data obtained within an independent psychological study. Lack of such data presents a limitation not only to our study but for other possible explorations of a similar kind. Moreover, even though reliability was observed when gathering the Dutch data, both similarity and typicality may still be regarded as considerably subjective phenomena, and hence, a human judgment of both similarity and typicality may suffer from additional forms of possible unreliability compared to when data is obtained by an ordinary physical measurement. The latter problem, however, represents an unavoidable aspect of experimentation with psychological data.

5. Conclusions

Our experiments comparing 62 similarity measures for binary data with human judgments of similarity via their ability to predict human assessment of typicality reveal several patterns and observations. Most importantly, human similarity results in overall very good predictions

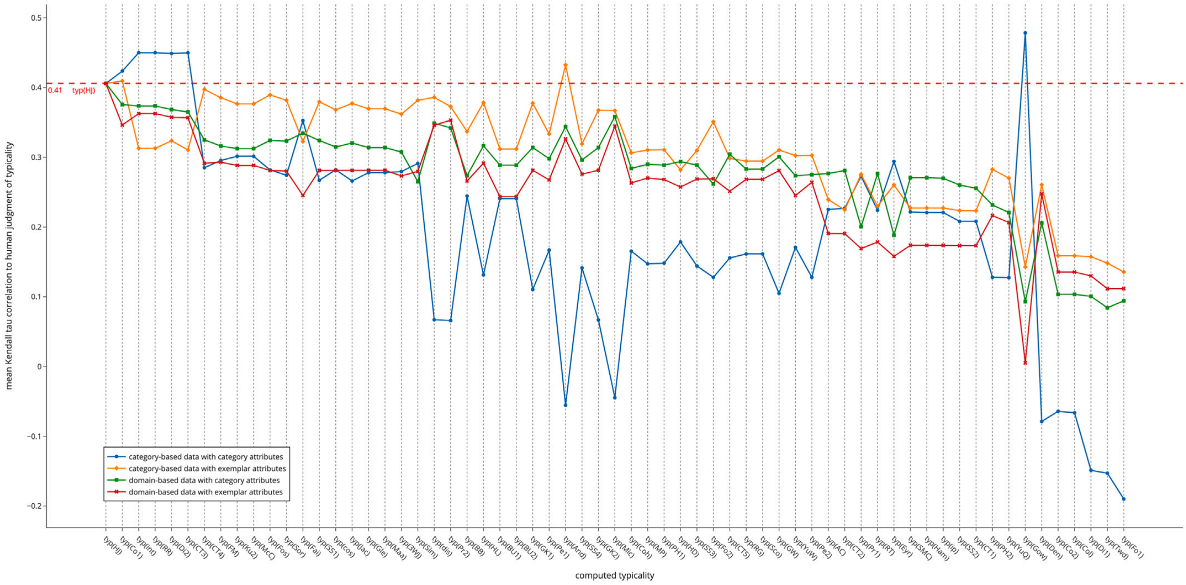

Fig. 9. Mean correlations of computed typicality to human judgment of typicality across categories of the animal domain.

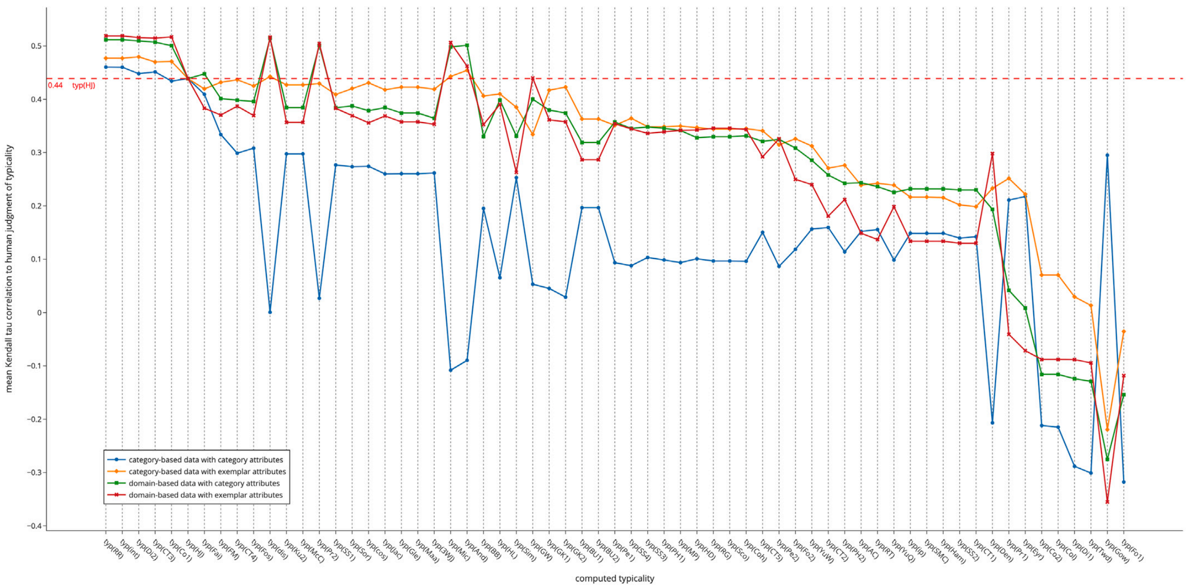

Fig. 10. Mean correlations of computed typicality to human judgment of typicality across categories of the artifact domain.

of typicality. For categories of the animal domain, it provides the best predictions. In this perspective, human similarity has a distinct place among the examined similarities as regards cognitive abilities.

On the other hand, human similarity ranks as the sixth best among all the explored similarities across all typicality predictions involved in our experiments. The experiments reveal a group of similarities, which includes human similarity, whose predictions of similarity are indeed strongly correlated with human assessment of typicality, as well as further observations worth further exploration.

As regards future research, we propose the following topics:

- The present experiments enable to compare similarities as regards their performance in a certain cognitive task (viz. prediction of typicality). A different experiment, however, should also be performed in which the existing similarity measures are compared as regards their ability to predict human judgment of similarity. This may reveal further, possibly different patterns and observations. The Dutch data, used in our experiment, allow for such kind of experiment.
- It became apparent that the quality of attributes which describe the exemplars plays a significant role in prediction of typicality of these exemplars. Since the quality of attributes is generally

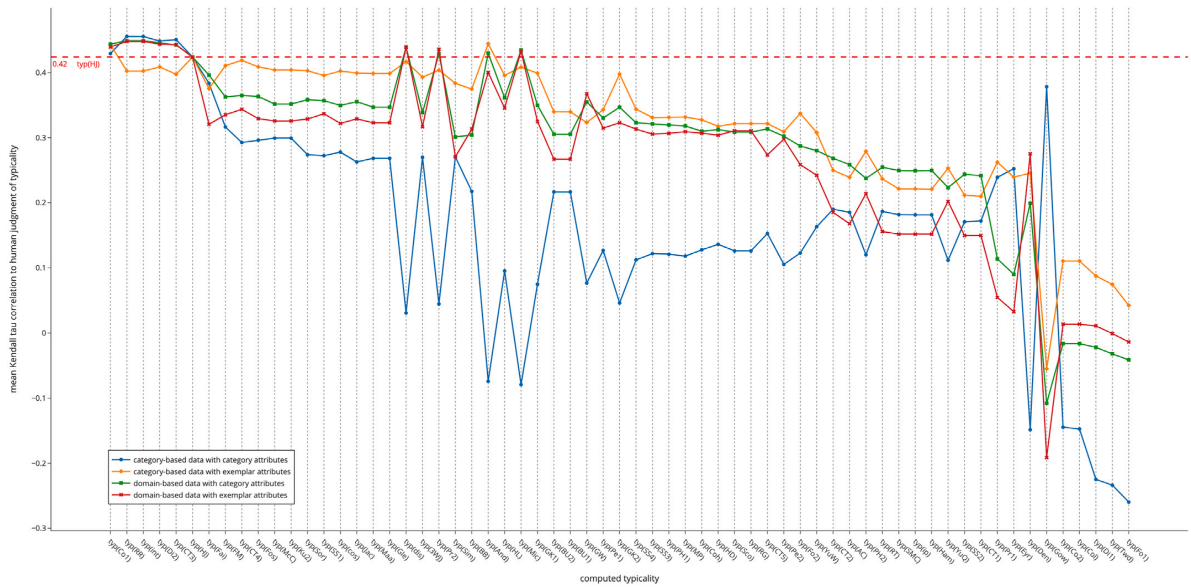

Fig. 11. Mean correlations of computed typicality to human judgment of typicality across all data.

regarded as important in a variety of cognitive tasks in the psychological literature, more focused studies shall be conducted in this direction. This includes possible quantitative measure of quality of a given set of attributes.

- In view of note 4.3, it seems to be of interest to compare the human assessment of similarity in the presence of context with the assessment with no context in the sense of note 4.3, as well as to perform a comparison with similarity degrees computed using similarity measures when binary attributes describing the exemplars are available. Such experiments may improve our understanding of the role of context for human assessment of similarity.
- It is apparent that for some categories (such as "mammal" in Fig. 2), the observed similarity measures differ in their capability to predict typicality to a larger extent compared to other categories (such as "fish" in Fig. 2). It seems of interest to explore in greater detail whether this phenomenon is due to the particular dataset used in our experiments or rather due to some general factor of psychological relevance.

CRediT authorship contribution statement

Radim Belohlavek: Formal analysis, Investigation, Methodology. **Tomas Mikula:** Formal analysis, Investigation, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Reference to the data is presented in the paper.

Acknowledgments

Supported partly by the project IGA 2023, reg. no. IGA_PrF_2023_026, of Palacký University Olomouc, Czech Republic.

Appendix. Similarity measures

The appendix presents 69 similarity measures for binary data we employ in the experiments along with additional information. In the formulas defining the similarity measures we denote for two binary vectors $x, y \in \{0,1\}^n$ by a, b, c, and d the numbers of attributes defined in Section 2.1. Hence, a, b, c, and d denote the number of attributes shared by x and y, possessed by x but not by y, possessed by y but not by x, and possessed neither by x nor by y, respectively. Thus,

$$n = a + b + c + d.$$

The measures are presented in Table 4. For each measure we include its abbreviation, its name (along with alternative names), a formula defining the measure, and a list of significant comparative papers in which this measure appears. The measures are ordered lexicographically by their abbreviations for ease of lookup. In our table, we refer to the following comparative papers, to which refer by the numbers 1–5 in the appendix:

1. Brusco, Cradit, and Steinley [1], which contains 71 similarity measures;
2. Choi, Cha, and Tappert, 2010 [2], which includes 60 similarity (and 16 dissimilarity) measures;
3. Hubálek, 1982 [4], which involves 20 similarity measures (in fact, it lists 43 measures from which 20 are selected after removing certain measures due to their equivalence with other involved measures or due to lack of required properties);
4. Todeschini, Consonni, Xiang, Holliday, Buscema, and Willett, 2012 [5], which employs 44 similarity measures (it includes 51 similarity measures, of which 7 were eliminated due to their equivalence with other measures);
5. Wijaya, Afendi, Batubara, Darusman, Altaf-Ul-Amin, and Kanaya, 2016 [6], which includes 62 similarity (and 17 dissimilarity) measures.

Remark 1. (a) Some similarity measures presented in Table 4 are not defined for certain values of a, b, c, and d, which naturally occur in data. The measures that suffer from this defect on our data are omitted in the graphs presenting results of our experiments in Section 4. In

Table 4
Similarity measures.

Symbol	Name	Formula	Source
AC	Austin-Colwell	$\frac{2}{\pi}\arcsin\sqrt{\frac{a+d}{n}}$	1, 3, 4
And	Anderberg	$\frac{\tau_1-\tau_2}{2n}$ with $\tau_1=\max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)$ $\tau_2=\max(a+c,b+d)+\max(a+b,c+d)$	1, 2, 5
BB	Braun-Blanquet	$\frac{a}{\max(a+b,a+c)}$	1, 2, 3, 4, 5
BU1	Baroni-Urbani-Buser 1	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	1, 2, 3, 4, 5
BU2	Baroni-Urbani-Buser 2	$\frac{\sqrt{ad}+a-b-c}{\sqrt{ad}+a+b+c}$	1, 2, 3, 4, 5
Coh	Cohen	$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	1, 4
Col	Cole	$\begin{cases}\frac{ad-bc}{(a+b)(b+d)} & \text{if } ad \geq bc\\ \frac{ad-bc}{(a+b)(a+c)} & \text{if } ad < bc \text{ and } d \geq a\\ \frac{ad-bc}{(b+d)(c+d)} & \text{otherwise}\end{cases}$	2, 3, 5
Co1	Cole (Cole 1)	$\frac{ad-bc}{(a+c)(c+d)}$	1, 4
Co2	Cole (Cole 2)	$\frac{ad-bc}{(a+b)(b+d)}$	1, 4
cos	cosine (Driver-Kroeber, Ochiai)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	1, 2, 4, 5
CT1	Consonni-Todeschini 1	$\frac{\ln(1+a+d)}{\ln(1+n)}$	1, 4
CT2	Consonni-Todeschini 2	$\frac{\ln(1+n)-\ln(1+b+c)}{\ln(1+n)}$	1, 4
CT3	Consonni-Todeschini 3	$\frac{\ln(1+a)}{\ln(1+n)}$	1, 4, 5
CT4	Consonni-Todeschini 4	$\frac{\ln(1+a)}{\ln(1+a+b+c)}$	1, 4, 5
CT5	Consonni-Todeschini 5	$\frac{\ln(1+ad)-\ln(1+bc)}{\ln(1+n^2/4)}$	1, 4, 5
Den	Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	1, 2, 4, 5
dis	dispersion	$\frac{ad-bc}{n^2}$	1, 2, 4, 5
Di1	Dice 1	$\frac{a}{a+b}$	1, 4
Di2	Dice 2	$\frac{a}{a+c}$	1, 4
Eyr	Eyraud	$\frac{n^2(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)}$	1, 2, 5
Fai	Faith	$\frac{a+0.5d}{n}$	1, 2, 4, 5
FM	Fager-McGowan	$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{\max(a+b,a+c)}}$	1, 2, 3, 5
Fos	Fossum	$\frac{n(a-\frac{1}{2})^2}{(a+b)(a+c)}$	1, 2, 4, 5
Fo1	Forbes 1	$\frac{na}{(a+b)(a+c)}$	1, 2, 3, 4, 5
Fo2	Forbes 2	$\frac{na-(a+b)(a+c)}{n\min(a+b,a+c)-(a+b)(a+c)}$	1, 2, 3, 5
Gle	Gleason (Dice, Sørensen, Czekanowski	$\frac{2a}{2a+b+c}$	1, 2, 3, 4, 5
GK1	Goodman-Kruskal 1	$\frac{\tau_1-\tau_2}{2n-\tau_2}$ with $\tau_1=\max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)$ $\tau_2=\max(a+c,b+d)+\max(a+b,c+d)$	1, 2, 5
GK2	Goodman-Kruskal 2	$\frac{2\min(a,d)-b-c}{2\min(a,d)+b+c}$	1, 4
Gow	Gower	$\frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	1, 2, 5
GW	Gilbert-Wells	$\ln\frac{n^3}{2\pi(a+b)(c+d)(a+c)(b+d)} + 2\ln\frac{n!a!b!c!d!}{(a+b)!(c+d)!(a+c)!(b+d)!}$	1, 2, 3, 5
Ham	Hamman	$\frac{a+d-b-c}{n}$	1, 2, 3, 4, 5
HD	Hawkins-Dotson	$\frac{1}{2}\left(\frac{a}{a+b+c}+\frac{d}{d+b+c}\right)$	1, 4
HL	Harris-Lahey	$\frac{a(2d+b+c)}{2(a+b+c)} + \frac{d(2a+b+c)}{2(b+c+d)}$	1, 4
int	intersection	a	2, 5
ip	inner product	$a+d$	2, 5
Jac	Jaccard (Jaccard-Tanimoto)	$\frac{a}{a+b+c}$	1, 2, 3, 4, 5
Ku1	Kulczynski 1	$\frac{a}{b+c}$	1, 2, 3, 5
Ku2	Kulczynski 2 (Driver-Kroeber)	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$	1, 2, 3, 4, 5
Maa	van der Maarel	$\frac{2a-b-c}{2a+b+c}$	1, 4
McC	McConnaughey	$\frac{a^2-bc}{(a+b)(a+c)}$	1, 2, 3, 4, 5
Mic	Michael	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	1, 2, 3, 4, 5
Mou	Mountford	$\frac{2a}{ab+ac+2bc}$	1, 2, 3, 4, 5
MP	Maxwell-Pilliner	$\frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	1, 4
Pe1	Pearson 1 (χ^2 statistical significance)	$\frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$	1, 2, 3, 5
Pe2	Pearson 2	$\sqrt{\frac{\chi^2}{n+\chi^2}}$ with χ^2 equal to Pe1	1, 2, 3, 5
Pe3	Pearson 3	$\sqrt{\frac{\rho}{n+\rho}}$ with ρ equal to PH1	2, 5
PH1	Pearson-Heron 1 (Phi)	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	1, 2, 3, 4, 5
PH2	Pearson-Heron 2	$\cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)$	2, 3, 5
Pr1	Peirce 1	$\frac{ad-bc}{(a+b)(c+d)}$	1, 4
Pr2	Peirce 2	$\frac{ad-bc}{(a+c)(b+d)}$	1, 3, 4
Pr3	Peirce 3	$\frac{ad+bc}{ab+2bc+cd}$	1, 2, 3, 5
RG	Rogot-Goldberg	$\frac{a}{2a+b+c}+\frac{d}{2d+b+c}$	1, 4
RR	Russel-Rao	$\frac{a}{n}$	1, 2, 3, 4, 5
RT	Rogers-Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	1, 2, 3, 4, 5
Sco	Scott	$\frac{4ad-(b+c)^2}{(2a+b+c)(2d+b+c)}$	1, 4
Sim	Simpson	$\frac{a}{\min(a+b,a+c)}$	1, 2, 3, 4, 5
SMC	simple matching coefficient (Sokal-Michener)	$\frac{a+d}{n}$	1, 2, 3, 4, 5
Sor	Sorgenfrei	$\frac{a^2}{(a+b)(a+c)}$	1, 2, 3, 4, 5
SS1	Sokal-Sneath 1	$\frac{a}{a+2b+2c}$	1, 2, 3, 4, 5
SS2	Sokal-Sneath 2	$\frac{2a+2d}{2a+b+c+2d}$	1, 2, 3, 4, 5
SS3	Sokal-Sneath 3	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d})$	1, 2, 3, 4, 5
SS4	Sokal-Sneath 4, Ochiai 2	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	1, 2, 3, 4, 5
SS5	Sokal-Sneath 5	$\frac{a+d}{b+c}$	1, 2, 3, 5
Sti	Stiles	$\log_{10}\frac{n(\lvert an-bc\rvert-\frac{1}{2}n)^2}{bc(n-b)(n-c)}$	1, 2, 5
Tar	Tarantula	$\frac{a(c+d)}{c(a+b)} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$	1, 2, 5
Twd	Tarwid	$\frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$	1, 2, 3, 5
YuQ	Yule (Yule Q)	$\frac{ad-bc}{ad+bc}$	1, 2, 3, 4, 5
YuW	Yule (Yule W)	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	1, 2, 3, 4, 5
3WJ	3W-Jaccard	$\frac{3a}{3a+b+c}$	1, 2, 4, 5

particular, there are 7 such measures: Ku1, Mou, Pe3, Pr3, SS5, Sti, Tar. For instance, the Mountfond measure given by $\frac{2a}{ab+ac+2bc}$ is not defined if $a = b = 0$ or $a = c = 0$.

(b) Some measures are not defined even for two objects which share the same attributes (i.e., for which $b = c = 0$), which is counterintuitive. One could redefine each such measure so that for $b = c = 0$ it yields its maximal value. We nevertheless refrained from this possible modification to obey the definitions presented in the literature.

Remark 2. We found a number of mistakes in the literature on similarity measures for binary data. In the following list, we include the significant ones pertaining to the measures we employ.

1. AC: 3 lists a slightly different formula for AC, namely $\frac{1}{50\pi}\sqrt{\frac{a+d}{n}}$, i.e., a formula yielding a value $100\times$ smaller than our formula.
2. Col: 2, 3, 5 list a different formula, namely $\frac{\sqrt{2}(ad-bc)}{\sqrt{(ad-bc)^2-(a+b)(a+c)(b+d)(c+d)}}$. This formula also appears in the original paper [18, p. 416] as a so-called mean square contingency, but is not meant as the similarity measure which the authors present in their paper. The Abydos library [19] lists our formula for Col.
3. Eyr: 3 lists a different formula, namely $\frac{a-(a+b)(a+c)}{(a+b)(a+c)(b+d)(c+d)}$.
4. FM: 1, 2, 3, and 5 list a different formula, $\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{\max(a+b,a+c)}{2}$, which is apparently wrong. Namely, the original paper [20] contains the formula we use as FM, and notes that this formula results by a modification of a formula used in [21].
5. Fos: 1 lists a different formula, $\frac{n(a-\frac{1}{2})^2}{\sqrt{(a+b)(a+c)}}$, which is an apparent misprint.
6. GW: 1, 2, 3, and 5 list a different formula, namely $\log a - \log n - \log(\frac{a+b}{n}) - \log(\frac{a+c}{n})$; 1 and 3 refer to [22], 2 does not contain a reference for this measure, and refers to 1 and 3. The original paper [22] includes our formula, as does [19].
Note also that 1, 2, 3, 4, and 5 list the so-called Johnson measure with a formula $\frac{a}{a+b} + \frac{a}{a+c}$. Clearly, this formula yields the value of $2 \cdot \mathrm{Ku}2$, hence we do not include the Johnson measure [23].
7. SS3: 2 contains a misprint in the formula for SS3.
8. Sti: 1, 2, and 5 list a different formula, namely $\log \frac{n(|ad-bc|-n/2)^2}{(a+b)(a+c)(b+d)(c+d)}$. Our formula comes from the original paper [24] and is also used in the Abydos library [19].

References

[1] M. Brusco, J.D. Cradit, D. Steinley, A comparison of 71 binary similarity coefficients: The effect of base rates, PLoS One 16 (2021) 4.

[2] S.S. Choi, S.H. Cha, C.C. Tappert, A survey of binary similarity and distance measures, J. Syst. Cybern. Inf. 8 (1) (2010) 43–48.

[3] J.C. Gower, P. Legendre, Metric and euclidean properties of dissimilarity coefficients, J. Classification 3 (1986) 5–48.

[4] Z. Hubálek, Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation, Biol. Rev. 57 (4) (1982) 669–689.

[5] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets, J. Chem. Inf. Model. 52 (11) (2012) 2884—2901.

[6] S.H. Wijaya, F.M. Afendi, I. Batubara, L.K. Darusman, M. Altaf-Ul-Amin, S. Kanaya, Finding an appropriate equation to measure similarity between binary vectors: Case studies on Indonesian and Japanese herbal medicines, BMC Bioinformatics 17 (1) (2016) 520.

[7] M.J. Warrens, Similarity Coefficients for Binary Data (Ph.D. thesis), Leiden University, The Netherlands, 2008, https://scholarlypublications.universiteitleiden.nl/access/item%3A2961378/view.

[8] G.L. Murphy, The Big Book of Concepts, MIT Press, Cambridge, Mass, 2002, http://dx.doi.org/10.7551/mitpress/1602.001.0001.

[9] R. Belohlavek, T. Mikula, Typicality: A formal concept analysis account, Internat. J. Approx. Reason. 142 (2022) 349–369.

[10] S. De Deyne, S. Verheyen, E. Ameel, W. Vanpaemel, M.J. Dry, W. Voorspoels, G. Storms, Exemplar by feature applicability matrices and other dutch normative data for semantic concepts, Behav. Res. Methods 40 (4) (2008) 1030–1048, http://dx.doi.org/10.3758/BRM.40.4.1030.

[11] E. Rosch, Principles of categorization, in: E. Rosch, B.B. Lloyd (Eds.), Cognition and Categorization, Erlbaum, Hillsdale, NJ, 1978, pp. 27–48.

[12] E. Rosch, C.B. Mervis, Family resemblances: Studies in the internal structure of categories, Cogn. Psychol. 7 (4) (1975) 573–605, http://dx.doi.org/10.1016/0010-0285(75)90024-9.

[13] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, P. Boyes-Braem, Basic objects in natural categories, Cogn. Psychol. 8 (3) (1976) 382–439, http://dx.doi.org/10.1016/0010-0285(76)90013-X.

[14] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2019, http://archive.ics.uci.edu/ml.

[15] W. Ruts, S. De Deyne, E. Ameel, W. Vanpaemel, T. Verbeemen, G. Storms, Dutch norm data for 13 semantic categories and 338 exemplars, Behav. Res. Methods Instrum. Comput. 36 (2004) 506–515.

[16] R. Belohlavek, T. Mikula, Dutch concepts. GitHub repository, 2023, https://github.com/mikulatomas/dutch-concepts.

[17] SciPy 1.0 Contributors, et al., Scipy 1.0: fundamental algorithms for scientific computing in python, Nature Methods (2020) http://dx.doi.org/10.1038/s41592-019-0686-2.

[18] L.C. Cole, The measurement of interspecific association, Ecology 30 (4) (1949) 411–424.

[19] Abydos: abydos.distance package. https://abydos.readthedocs.io/en/latest/abydos.distance.html.

[20] E.W. Fager, J.A. McGowan, Zooplankton species groups in the north pacific, Science 140 (3566) (1963) 453–460.

[21] E.W. Fager, Determination and analysis of recurrent groups, Ecology 38 (4) (1957) 586—595.

[22] N. Gilbert, T.C.E. Wells, Analysis of quadrat data, J. Ecol. 54 (3) (1966) 675–685.

[23] S.C. Johnson, Hierarchical clustering schemes, Psychometrika 32 (3) (1967) 241–254.

[24] E.H. Stiles, The association factor in information retrieval, J. ACM 8 (2) (1961) 271–279.

Appendix C

Schemes to compute typicality: Revisiting Rosch and Mervis' approach

This preprint (Belohlavek and Mikula, 2024c) proposes an extended version of Rosch and Mervis' scheme and shows their relationship to the similarity-based scheme. We also describe new attribute weight based on characteristicness, outperforming the previously tested typicality schemes. Theoretical foundations are described in Section 3.2. The results of these experiments are briefly described in Section 4.3. The preprint results from joint research with my supervisor Radim Bělohlávek.

Schemes to compute typicality: Revisiting Rosch and Mervis's approach

Radim Belohlavek, Tomas Mikula

Palacký University Olomouc, Czech Republic

(Preprint to be submitted)

Abstract

We revisit Rosch and Mervis' approach to typicality of exemplars in human categories. We argue that their definition naturally leads to two rather different schemes to compute degrees of typicality. While one was proposed in Rosch and Mervis' seminal paper, the other involves a measure of similarity. Experimental evaluation utilizing a large-scale data regarding human categories reveals a high correlation of both schemes with a human judgment of typicality. We nevertheless examine their relationship and prove that Roch and Mervis' typicality formula, as well as its natural generalization, which takes into account the absence of attributes in a category rather than restricting to presence like the original formula, are equivalent to a particular case of the similarity-based formula with a properly chosen similarity measure. These considerations lead to a technically simple but conceptually significant extension of Rosch and Mervis' formula that involves the concept of attribute characteristicness. We provide an extensive experimental evaluation of the considered typicality formulas and establish that the new typicality formula outperforms the original Rosch and Merivis' formula, the variants, as well as the similarity-based scheme. Our approach hence provides a novel psychological account of typicality which is more effective than the previous ones in terms of agreement with human assessment of typicality.

Keywords: concept, typicality, similarity, psychology of concepts

Email addresses: `radim.belohlavek@acm.org` (Radim Belohlavek), `mail@tomasmikula.cz` (Tomas Mikula)

1. Introduction

1.1. Typicality in the psychology of concepts

Typicality of exemplars in human categories represents a widely studied notion in the psychology of concepts [17]. The interest in typicality goes back to the various explorations in the mid-1970s in the internal structure of concepts, most importantly those led by Eleanor Rosch. The revealed fundamental limitations of the classical view of concepts according to which a concept is determined by a set of yes/no (bivalent, binary) conditions (attributes, features) which are necessary and jointly sufficient. That is, an object is covered by (or, is a member of) the concept (or category in terms commonly used in the psychology of concepts) if and only if the object satisfies each of these conditions. In the mid-1970s, however, it became apparent that concepts have a graded structure: Various phenomena had experimentally been found to be a matter of degree rather than bivalent (yes/no). In addition, important phenomena had been observed that were not accounted for by the classical view. Typicality, which is discussed in the first findings by Rosch et al. [18, 19, 20], represents such a phenomenon.

The classical view does not account for typicality, at least not directly, which represents a considerable shortcoming. Namely, according to the classical view, all members of a category have an equal status with respect to the category. On the other hand, people naturally regard some objects more typical of a given category than other objects. Further research has shown that people are even capable of assigning degrees of typicality (also called typicality ratings) to objects for a given category in a consistent manner. Importantly, typicality has been found to have a high cognitive significance; see, e.g., [2, 17, 19]. For one, people tend to agree on typicality ratings. Moreover, typicality is reported to predict performance in a variety of cognitive tasks, including learning of categories (typical objects are learned more quickly), deciding membership in categories (decisions on typical objects are quicker), and production of category exemplars (typical exemplars are generated first). Typical items are also useful in making inferences about categories and serve as so-called cognitive reference points. Due to the high significance of typicality, the phenomenon continues to be a subject of vivid psychological research; see, e.g., [10, 32].

1.2. Explanations of typicality

In their seminal paper [19], Rosch and Mervis put forward a hypothesis of what makes an object typical in a category. This hypothesis was confirmed by experiments by the authors [19] and had later been examined by numerous other studies; see, e.g., the monograph [17], in which typicality occupies a significant part. The hypothesis is based on the notion of sharing of attributes by members of the given category and is described as follows [19, p. 575]:

> ... members of a category come to be viewed as prototypical of the category as a whole in proportion to the extent to which they bear a family resemblance to (have attributes that overlap those of) other members of the category.

In addition, several other possible explanations of typicality of an item have been suggested and tested in later studies, including similarity to central tendency (central tendency being, e.g., the average of a numerical characteristic of an item), closeness to ideals in goal-oriented categories (ideals represent characteristics that items should possess if they are to serve the goal associated with a category), frequency of instantiation (i.e. frequency of encounter with the item as a member of a given category), and familiarity (i.e., frequency of encounter across all contexts); see, e.g., [2, 16, 17] and also [12]. A more recent research also emphasizes the role of context (situation) in which typicality is assessed [32]. The resulting instability of typicality resulting from dependence on context even led the authors in [10] to distinguish between the so-called structural typicality (representing stability) and functional typicality (representing context-dependence and thus instability).

In spite of several alternative hypotheses, the family resemblance hypothesis of Rosch and Mervis [19] mentioned above appears to remain the simplest and most commonly accepted explanation of typicality. It is due to this fact that Rosch and Mervis' explanation forms the basis of our approach.

1.3. Our contribution

We revisit Rosch and Mervis' approach to typicality [19] from a theoretical as well as an experimental perspective. While Rosch and Mervis propose a particular formula that is derived as a computational counterpart of their verbal characterization of typicality, we argue that their characterization allows for a different computational interpretation, namely, one that explicitly involves a measure of exemplars' similarity.

For the purpose of analyzing the two computational interpretations of the characterization of typicality, we extend the original Rosch and Mervis' formula to a scheme which substantially generalizes the formula, yet preserves its meaning. Unlike the original formula, which only involves the presence of attributes in exemplars, our proposed scheme also involves attributes' absence and, moreover, allows to set the significance of attributes' presence and absence using coefficients corresponding to presence and absence, respectively. In this setting, the original Rosch and Mervis' formula becomes a particular case in which the significance coefficient for absence is set to 0. As regards the comparison of the generalized scheme subsuming the Rosch and Mervis' formula with the new and technically rather different scheme involving a measure of exemplars' similarity, we prove that Rosch and Mervis' formula, and even the generalized scheme, is essentially equivalent to a particular case of the similarity-based scheme. In particular, it yields—up to a certain scaling function—the same assessment of typicality as a particular case of the similarity-based scheme with the involved measure of similarity set accordingly.

We, nevertheless, reconsider the scheme inspired by the Rosch and Mervis' formula and suggest its conceptual extension. While the original Rosch and Mervis' formula involves a weight for every attribute representing to which this attribute is shared by the category members, we propose to consider a more general prospect of attributes' weights, namely one that takes into account the outside of a category rather than the category's interior only to which Rosch and Mervis' formula implicitly restricts. The motivation consists in that research on typicality suggests that in addition to how an exemplar relates to the category's interior, typicality of this exemplar in the category also depends on its relationship with the outside of the category. While this suggestion is mentioned in the literature, including the seminal paper by Rosch and Mervis [19, p. 575], it is ignored by Rosch and Mervis' formula. We propose a particular notion of an attribute weight, which may be interpreted as attribute characteristicness and which is analogous to some notions considered in the literature, and a novel formula to compute typicality that exploits this notion of weight.

All the three considered schemes to compute typicality are subject to an experimental evaluation utilizing the so-called Dutch data—the now available extensive psychological data involving several common language categories, tens of exemplars in these categories, and a large number of attributes describing these exemplars, as well as other collected data including a human judgment of typicality. The scope of the presented evaluation significantly exceeds the one of the few previous quantitative evaluations, which were performed on a data of a rather limited size and quality. The experiments reveal that the three prospects to compute typicality indeed yield a considerable agreement with a human judgment. Moreover, while the performance of the scheme based on the original Rosch and Mervis' is comparable to that of the similarity-based scheme, both of these schemes are generally outperformed by the third one utilizing attribute characteristicness.

To sum up, we provide a considerably deeper insight into the Rosch and Mervis' formula to compute degrees of typicality and demonstrate its empirical validity. We also suggest an alternative, technically rather different computational scheme explicitly based on similarity, which is proven to subsume Rosch and Mervis' formula. Most importantly, we reconsider the scheme behind Rosch and Mervis' formula and obtain a novel formula utilizing a new prospect of attribute weight which considers the outside of a category. As regards agreement with a human judgment, this last formula outperforms the previous ones. In addition to resulting in a practical and well-performing formula to compute degrees of typicality, our paper confirms that both the relationship of an object to the interior and the outside of a category are significant in determination of the object's typicality.

2. Schemes to compute typicality and their relationship

In accordance with common practice, we assume that the considered category is represented by (and may thus be identified with) a subset A of a universe set X of the considered objects. If, for instance, X consists of the objects in the animal domain, then category "bird" is represented by the set of all exemplars of birds contained in X. Moreover, we assume that there is a set Y of binary attributes (i.e., yes/no attributes) and an incidence relationship I between the objects in X and the attributes in Y, with xIy indicating that the object x (e.g., $x =$ sparrow) has the attribute y (e.g., $y =$ can fly), and $x\overline{I}y$ indicating the opposite. Information of this kind is commonly represented by a binary matrix, in which the rows and columns correspond to objects in X and attributes in Y, respectively, and the incidence relation I is represented by the matrix values 0 and 1.

2.1. Rosch and Mervis' formula

Rosch and Mervis [19] proposed a simple formula to compute a degree of typicality which involves a particular notion of attribute weight. The definition is supposed to formalize their verbal characterization of typicality quoted in section 1.2. In particular, they define the weight $w(y, A)$ of an attribute y in Y with respect to a category A by

$$w(y, A) = |\{x\,;\, x \text{ is in } A \text{ and has } y\}|, \tag{1}$$

i.e., $w(y, A)$ is the number of objects of the category A sharing the attribute y. The degree $typ_{\mathrm{RM}}(x, A)$ of typicality of an object x in A is then defined by

$$typ_{\mathrm{RM}}(x, A) = \sum_{y \in Y,\ xIy} w(y, A), \tag{2}$$

i.e., as the sum of weights of all the attributes y possessed by x. The definition is present in a verbal form as part of Experiment 1 in [19, p. 580] in which the authors tested agreement of the formula with a human assessment of typicality using what may considered a rather small data given in view of the data we use in our study.

2.2. A scheme based on a measure of similarity

Even though formula (2) by Rosch and Mervis is inspired by hypothesis quoted in section 1.2, it is argued in [3] that more directly, the hypothesis leads to a technically rather different scheme to compute typicality, which the authors present in [3] in the context of typicality in formal concept analysis. In this section, we present a natural generalization of this scheme that applies to arbitrary categories A, i.e., subsets A of a universe set X of the considered objects, rather than to the so-called formal concepts only as in [3]. Since Rosch and Mervis' hypothesis explicitly refers to similarity of an object to other objects in the category, it is natural to consider a scheme that employs a function

$$sim : X \times X \to [0, 1]$$

assigning to every two objects $x_1, x_2 \in X$ a number $sim(x_1, x_2) \in [0, 1]$ interpreted as a degree to which x_1 and x_2 are similar. Similarity of x to the objects x_1 in A, which underlies Rosch and Mervis' view of typicality, may then be interpreted as the average similarity of x to all the objects $x_1 \in A$. This leads to the following definition of the degree $typ_{sim}(x, A)$ of typicality of x in A:[1]

$$typ_{sim}(x, A) = \frac{\sum_{x_1 \in A} sim(x, x_1)}{|A|}. \tag{3}$$

This way, typicality is parameterized by a similarity function *sim*. Since each object x is described by a binary vector, whose components correspond to the attributes in Y and represent whether x has a particular attribute y, the similarity function may be chosen from over fifty commonly recognized measures of similarity for binary data, such as the well-known Jaccard measure, the simple matching coefficient, and others; see, e.g., [7, 9, 14, 27, 30, 29]. Some of them shall be employed in our experimental evaluation below.

2.3. A scheme extending Rosch and Mervis' formula

Before we analyze the relationship between Rosch and Mervis' typicality formula of section 2.1 and the similarity-based scheme of section 2.2, let us point out a natural extension of Rosch and Mervis' formula. It consists in realizing that the

[1]Average similarity is mentioned in some psychological studies; see, e.g., [2, p. 630]. We use $[0, 1]$ for the range (i.e. similarity is scaled), but $\mathbb{R}^+$ is also a natural option (non-scaled).

formula only focuses on shared presence of a given attribute by members of the category and disregards shared absence. To symmetrize the approach, we propose to consider two kinds of attribute weight: The Rosch and Mervis' weight (1), now denoted $w^+(y, A)$, and the symmetric weight $w^-(y, A)$ representing the number of objects of A that do not have the attribute y, i.e., we consider

$$\begin{aligned} w^+(y, A) &= |\{x\,; x \text{ is in } A \text{ and has } y\}|, \text{ and} \\ w^-(y, A) &= |\{x\,; x \text{ is in } A \text{ and does not have } y\}|. \end{aligned}$$

The corresponding degree of typicality of x in A is then defined by the following presence/absence extension of Rosch and Mervis' formula:

$$typ_{\mathrm{RM}^\pm}(x, A) = \sum_{y \in Y,\ xIy} w^+(y, A) + \sum_{y \in Y,\ x\bar{I}y} w^-(y, A) \tag{4}$$

Intuitively, this formula says that typicality of an object in a category is proportional to the extent to which the object possesses attributes commonly shared by members of the category and does not have attributes commonly absent on the category members. Moreover, as is known from the studies in similarity, people usually regard attribute presence as more significant than absence. We hence propose to use non-negative weights a^+ and a^- that allow to set the significance of shared presences and shared absences differently, and define a weighted presence/absence scheme by

$$typ_{\mathrm{RM}^\pm}^{a^+,a^-}(x, A) = a^+ \cdot \sum_{y \in Y,\ xIy} w^+(y, A) + a^- \cdot \sum_{y \in Y,\ x\bar{I}y} w^-(y, A). \tag{5}$$

It is clear that for $a^+ = 1$ and $a^- = 0$, formula (5) yields the original Rosch and Mervis' typicality formula, and for $a^+ = 1$ and $a^- = 1$ it becomes formula (4).

2.4. Relationships between the schemes

The original Rosch and Mervis' formula typ_{RM}, as well as its extensions $typ_{\mathrm{RM}^\pm}$ and $typ_{\mathrm{RM}^\pm}^{a^+,a^-}$, are technically rather different from the similarity-based scheme typ_{sim}. Yet, the following theorem, whose proof is provided in the appendix, and its corollaries show that all the three presence/absence schemes, in fact, result by a particular scaling from the similarity-based scheme with an appropriate choice of similarity measure *sim*.

For our purpose, we denote by $\{x\}^\uparrow$ the set of all attributes shared by the object x, i.e.,

$$\{x\}^\uparrow = \{y\,; x \text{ has } y\},$$

and by $|S|$ the number of elements in a set S. We consider the similarity function

$$\mathrm{SMC}^{a^+,a^-}(x_1, x_2) = \frac{a^+ \cdot |\{x_1\}^\uparrow \cap \{x_2\}^\uparrow| + a^- \cdot |Y - (\{x_1\}^\uparrow \cup \{x_2\}^\uparrow)|}{|Y|} \tag{6}$$

parameterized by non-negative reals a^+ and a^-. In (6), the factor $|\{x_1\}^\uparrow \cap \{x_2\}^\uparrow|$ represents the number attributes shared by the objects x_1 and x_2, and $|Y - (\{x_1\}^\uparrow \cup \{x_2\}^\uparrow)|$ is the number of attributes absent on both x_1 and x_2. SMC^{a^+,a^-} hence represents a reasonable similarity function (see below for particular cases).

Theorem 1. *For arbitrary $a^+, a^- \geq 0$, each object x, and any category A,*

$$typ^{a^+,a^-}_{RM^\pm}(x, A) = |A| \cdot |Y| \cdot typ_{\mathrm{SMC}^{a^+,a^-}}(x, A)$$

where $typ_{\mathrm{SMC}^{a^+,a^-}}(x, A)$ is determined by SMC^{a^+,a^-} according to (3).

Before discussing implications of theorem 1, let us consider two corollaries, for which purpose we consider two particular choices of a^+ and a^-.

(a) $a^+ = 1$ and $a^- = 0$: In this case, the similarity function in (6) shall be denoted RR, i.e.,

$$\mathrm{RR}(x_1, x_2) = \mathrm{SMC}^{1,0}(x_1, x_2) = \frac{|\{x_1\}^\uparrow \cap \{x_2\}^\uparrow|}{|Y|}.$$

The function RR is, in fact, one of the existing similarity measures, called the Russel-Rao measure [21].

(b) $a^+ = 1$ and $a^- = 1$: In this case, the similarity in (6) shall be denoted SMC, i.e.,

$$\mathrm{SMC}(x_1, x_2) = \mathrm{SMC}^{1,1}(x_1, x_2) = \frac{|\{x_1\}^\uparrow \cap \{x_2\}^\uparrow| + |Y - (\{x_1\}^\uparrow \cup \{x_2\}^\uparrow)|}{|Y|}.$$

This function is one of the best known similarity measures and is known as the simple matching coefficient or the Sokal-Michener measure [23].

The following corollaries of theorem 1 show that the long-established similarity measures RR and SMC are exactly the measures corresponding to the original Rosch and Mervis' formula typ_{RM} and its presence/absence extension $typ_{\mathrm{RM}^\pm}$. The first concerns Rosch and Mervis' formula (this relationship was considered in [3]):

Corollary 1. *For each object x and an arbitrary category A,*

$$typ_{\mathrm{RM}}(x, A) = |A| \cdot |Y| \cdot typ_{\mathrm{RR}}(x, A)$$

where typ_{RR} is the typicality (3) induced by the Russell-Rao measure.

The second one concerns the extension of Rosch and Mervis' formula incorporating absence of attributes:

Corollary 2. *For each object x and an arbitrary category A,*

$$typ_{RM^{\pm}}(x, A) = |A| \cdot |Y| \cdot typ_{\mathrm{SMC}}(x, A)$$

where typ_{SMC} is the typicality (3) induced by the simple matching coefficient.

As to the implications of the above theorem and its corollaries, they firstly provide a better understanding of the original Rosch and Mervis' typicality formula vis-à-vis their hypothesis on typicality as being rooted in similarity (resemblance; see section 1.2). In particular, corollary 1 reveals that while Rosch and Mervis' formula does not explicitly involve similarity, it does so implicitly as it is equivalent via a simple scaling to a general similarity-based formula for typicality with a particular choice of the Russell-Rao similarity. A technical consequence which is worth mentioning is that both the Rosch and Mervis' formula typ_{RM} and the Russell-Rao-similarity-based formula typ_{RR} results in the same ordering of objects by their computed degrees of typicality. Analogous remarks apply to the natural extension $typ_{\mathrm{RM}^{\pm}}$ of Rosch and Mervis' formula that takes shared absence of attributes into account in the assessment of typicality, and to the generalization of this natural extension that allows for setting the significance of shared presences and shared absences differently.

As to the converse relationship, it is clear that the general similarity-based scheme to compute typicality, typ_{sim}, is more general than the presence/absence-based extended Rosch and Mervis' scheme with weights, $typ_{\mathrm{RM}^{\pm}}^{a^+,a^-}$, due to the great variety of similarity measures. Nevertheless, Rosch and Mervis' formula when looked at from a general perspective, leads to a new scheme to compute typicality which shall be examined in the next section.

2.5. New scheme utilizing the outside of a category

2.5.1. Beyond the essential restriction of Rosch and Mervis' formula

Rosch and Mervis' formula involves a significant restriction not apparent at the first sight. It derives from the nature of a weight $w(y, A)$ of the attribute y w.r.t. the category A and consists in that the definition of the weight by (1) only utilizes information inside the category in that it only on to which objects *in the category* the attribute y applies. As a result, the weights $w(y, A)$ ignore the outside of the category.

The original role of the weight $w(y, A)$, as proposed by Rosch and Mervis, is to express the presence of the attribute on the category members. In a more general sense, though, we now propose to utilize the role of an attribute weight $w(y, A)$ in the typicality formula (2) is to express a significance of the weight for a determination of typicality of the objects in the category. In view of such general purpose, ignoring the category's outside appears overly restrictive. In fact, Rosch and Mervis' verbal characterization of typicality refers to the category's outside, but their formula only utilizes the category's inside.

We hence propose to follow the more general understanding of attribute weights mentioned in the previous paragraph and utilize the category's outside for this purpose. As we demonstrate by our experimental evaluation below, this new approach results in a considerably better agreement with a human judgment of typicality compared to the previous approaches.

Note at this point that Rosch and Mervis also mention [19, p. 575] a converse view of typicality according to which the objects typical for a given category shall bear least family resemblance to other categories, and in this manner also refer to the outside of a given category. They present in their Experiment 2 an experimental evaluation of this converse characterization in which they involve, as the "other categories," what they refer to "categories at the same level of linguistic contrast." From our perspective, this represents an interesting proposition worth further exploration. Nevertheless, we do not examine this converse characterization in our paper because it requires availability of the contrast categories, as well as some further information involved in Rosch and Mervis' Experiment 2, and, in this sense, it is qualitatively different from the characterization we focus on.

2.5.2. Attribute characteristicness as a weight

The particular form of attribute weight described in the previous paragraph derives from a natural idea of a weight representing what we call the characteristicness of an attribute. Note that similar ideas have appeared in the literature, although mostly in a verbal, non-formalized manner, under various names, including distinctiveness [13, 6], centrality [24, 25, 28, 1], diagnosticity [15, 8], and typicality [31]; see also [17].

Put briefly, we consider an attribute y characteristic of a category A to the extent to which being a member of A is roughly equivalent to having y. This idea offers two possible ways to formalize it. Let us first denote by $\{y\}^{\downarrow}$ the set of all objects shared by the attribute y, i.e.,

$$\{y\}^{\downarrow} = \{x\,;\, x \text{ has } y\}.$$

The first way leads to the formula

$$w(y, A) = \frac{|\{y\}^{\downarrow} \cap A|}{|A|} \cdot \frac{|\{y\}^{\downarrow} \cap A|}{|\{y\}^{\downarrow}|}. \tag{7}$$

The first factor, $\frac{|\{y\}^{\downarrow} \cap A|}{|A|}$, is naturally interpreted as the truth degree (or, extent) to which the objects in A have y; the second factor, $\frac{|\{y\}^{\downarrow} \cap A|}{|\{y\}^{\downarrow}|}$, may be regarded as the truth degree to which the objects sharing y belong to A. The multiplication $\cdot$ of these two factors represents a many-valued conjunction. Hence, the weight $w(y, A)$ represents the truth degree of the assertion "the object sharing y belong to A and the objects in A share y." Note that instead of multiplication, which is known as the Goguen conjunction in the field of many-valued logic, one can use another many-valued conjunction, but we employ the multiplication.

The second way leads to

$$w(y, A) = \frac{|\{y\}^{\downarrow} \cap A|}{|A|} \cdot \frac{|(X - A) - \{y\}^{\downarrow}|}{|X - A|}. \tag{8}$$

In this case, the second factor, $\frac{|(X-A)-\{y\}^{\downarrow}|}{|X-A|}$, is interpreted as the degree to which the objects outside of A do not have y. Again, the multiplication represents a many-valued conjunction.

A difference between (7) and (8) consists in that in (8), each object is considered only once, the first time in A (first factor), the second time in $X - A$ (second factor), while in (7), the objects that belong to both A and $\{y\}^{\downarrow}$ are considered twice (in both factors). In our experiments, we use the first formula, (7) because it yields a slightly better agreement with a human assessment of typicality.

2.5.3. New formula for typicality

In view of the considerations above, we now propose the formula

$$typ_w(x, A) = \sum_{y \in \{x\}^{\uparrow}} w(y, A), \tag{9}$$

with the attribute weight $w(y, A)$ defined as the characteristicness of y with respect to A by (7).

3. Experiments

3.1. Data

For our purpose, the Dutch data [11] is uniquely suitable for our purpose, as it provides perhaps the most comprehensive data regarding common human categories and their numerous characteristics, including typicality and similarity. In this section, we provide a brief description of the data, particularly the parts we use, and our comments regarding usability in experiments along with our technical modifications in this regard. The Dutch data has been gathered by psychologists at the University of Leuven in a thorough, carefully designed study involving hundreds of human respondents. It basically provides information regarding common language concepts (categories), binary attributes (features) relevant to these categories, objects (exemplars) in these categories, and various psychologically relevant characteristics.

In particular, the data involves 16 linguistic categories. These include both the so-called natural kind and artifact categories, as these two kinds are commonly believed to have distinct properties. Each category is represented by a number of objects (exemplars), such as a robin for the category "bird." There are 10 natural kind categories: "fruit" (30 exemplars); "vegetables" (30); "professions" (30); "sports" (30); the animal categories "amphibians"(5),[2] "birds" (30), "fish" (23), "insects" (26), "mammals" (30), and "reptiles" (22).[3] In addition, there are 6 artifact categories: "clothing" (29), "kitchen utensils" (33), "musical instruments" (27), "tools" (30), "vehicles" (30), and "weapons" (20).[4]

These categories comprise 249 exemplars for the natural kind and 166 exemplars for the artifact categories, which were obtained from humans and are representative of the respective categories.[5] Coverage by these categories is considerable; for instance, the animal categories cover a rather large part of the known animal domain. The objects (exemplars) and attributes (features) were obtained by processes described in

[2]Since the category "amphibians" only contains 5 exemplars, and since these exemplars are included in the category "reptiles," we omit it in most of our considerations below; see [11] for reasons to include the exemplars of "amphibians" in "reptiles."

[3]The exemplar-by-feature applicability matrices, which we describe below and use in our experiments, contain only 20 exemplars of the category
"reptiles," because the respondents who were to fill in these matrices turned out to not to be familiar with two exemplars (komodo and iguanodon). We hence exclude these two exemplars from our experiments.

[4]Here, we use the plural in category names, as the authors do [11]; below, we use singular, i.e., "bird" rather than "birds" to be consistent with our previous writings.

[5]In addition to the 5 amphibians included in reptiles and two omitted exemplars of reptiles (see above), note that three exemplars of artifact categories are included in two distinct categories.

domain	objects	category attributes	exemplar attributes
animal	129	225	764
artifact	166	301	1,295

Table 1: Domain-based binary matrices used in our experiments.

[11]. In particular, the attributes were generated by 1 003 respondents in two ways: First, respondents were asked to list relevant attributes for a given category (these are called category attributes). Second, they were asked to list relevant attributes for each object involved in the data (these are called exemplar attributes). Furthermore, unions of all exemplar features listed for all the objects in a given category were considered, as well as the union of all exemplar features of all the objects in the animal domain, and an analogous union of exemplar features for the artifact domain.

An essential part of the data are the so-called exemplar-by-feature applicability matrices. These are various matrices in which the rows and columns correspond to some of the objects and attributes, respectively, and the entries contain information about whether a particular object has or does not have a particular attribute. Each of the matrices was filled separately by four respondents. The data also contains the corresponding aggregated matrices, in which the values, viz. 0, 1, 2, 3, and 4, indicate the number of respondents who agreed on that the respective object has the respective attribute. To obtain binary matrices (and thus data with binary attributes) from these aggregated matrices, one naturally thresholds the matrix entries. We present our experiments for a threshold equal to 2. Hence, our binary matrices contain 1 in the entry corresponding to the object x and the attribute y if at least two respondents agreed that x has y.

In particular, we use the binary matrices described in table 1. For instance, the 129×225 binary matrix referred to by the first row in table 1 describes which of the 129 objects in the animal domain have which of the corresponding 225 category attributes; the 129 objects are all the objects of the categories "amphibian", "bird", "fish", "insect", "mammal", and "reptile", and the 225 category attributes are all attributes listed as category attributes for these six categories. Likewise, the 129×764 matrix describes which of the objects in the animal domain have which of the corresponding 764 exemplar attributes, i.e., all the attributes listed as exemplar attributes for some of the 129 exemplars in the animal domain.

Typicality ratings, which are present in the Dutch data, were obtained from 112 respondents. For each of the 16 categories and each object in the respective category, the data contains a typicality rating on the scale of 1 (very atypical) to 20 (very typical).

The pairwise similarity ratings of the Dutch data come partly from the previous study [22], in which the ratings were obtained for ten of the present categories from 42 participants. The ratings for the other categories were obtained from 92 respondents in [11], who also provided additional ratings for the ten categories involved in [22] to improve reliability. For every category—except for "amphibians," whose five exemplars are included in "reptiles"—and each pair of objects, the data contains a similarity rating on the scale 1 (totally dissimilar) and 20 (totally similar).

Since the original data contains some minor semantic and technical faults, as well as inconveniences as regards a possible machine processing of the data, we modified the data as follows. For one, since the original data contains some wrongly formatted comma-separated files, we transformed them into a valid format. In addition, the names of some objects and attributes are spelled differently across multiple files in the original data; we therefore unified these names. We also converted all names to lowercase to unify them. No changes were made to the data itself. The result is easily machine-processable data. The corrected version of Dutch data, along with a convenient Python wrapper, is publicly available on GitHub [4].

3.2. Rationale of our experiments

We conducted a set of experiments to test the plausibility of the newly proposed scheme of typicality typ_w. We do so by comparing calculated degrees of typicality with human typicality ratings available in the Dutch data. To evaluate the performance not only with respect to the so-called ground truth but also to the similarity-based scheme, we include the degrees of typicality computed by the similarity-based formulas typ_{sim} corresponding to ten similarity measures sim. These similarity measures were selected according to their performance in a previous study, which involved a total of 69 similarity measures [5]. Importantly, the selected similarity-based formulas include typ_{HJ}, i.e., the formula based on the similarity HJ provided by human respondents which is available in the Dutch data.

As explained in section 3.1, the Dutch data contains data describing attributes of the objects (exemplars) of multiple categories from the animal and artifact domains. These data enable one to compute the weight of attribute $w(y, A)$ with respect to the given set of objects A from the given category (e.g., "bird") by formula (7). For illustration, table 2 includes the top 10 most characteristic attributes and the bottom 10 least characteristic attributes for the exemplar "sparrow" in the category "bird." Note that attributes with weight equal to 1.0 (e.g., "has two paws", "has feathers") are included only in the exemplars from the category "bird" – they are the most characteristic ones. On the other hand, more general attributes have a value close to zero (e.g., "is nice", "is brown"). The degrees of characteristicness for the particular attributes are hence in an intuitive agreement with a human view.

attribute y	$w(y, A)$
has two paws	1.0000
is a bird	1.0000
has feathers	1.0000
has air sacs	1.0000
has a bill	1.0000
has two wings	0.9677
has a beak	0.8824
eats seed	0.7333
builds nests	0.7143
eats worms	0.6877
...	...
is beautiful	0.1593
lives in warm countries	0.1469
herds	0.1337
is found in the garden	0.1280
flutters	0.1067
its excrements are found on the street	0.1042
is a collective noun	0.0955
is nice	0.0762
is brown	0.0672
dark colour	0.0629

Table 2: Top/bottom 10 attributes of "sparrow" exemplar from category "bird" with category attributes according to their characteristicness.

According to formula (9), the typicality of an object x from the given category is calculated as the mean value of weights of all attributes shared by x. Table 3 provides an example from category "bird", which includes three calculated degrees of typicality: typ_w, which is based on the attribute characteristicness $w(y, A)$; the similarity-based typicality typ_{RR} based on the Russel-Rao similarity, which is—according to corollary 1—equivalent to the original Rosch and Mervis formula typ_{RM}; and typ_{HJ} which is based on the human similarity ratings.[6] Each of the typicality

[6]The remaining eight typicality formulas typ_{sim}, i.e., those based on the other eight measures

rating columns is accompanied by a column that lists exemplars ordered from the most typical to the least typical one according to the computed degrees of typicality.

Since the Dutch data contains typicality ratings gathered from human respondents, we compare these ratings with the computed degrees of typicality. For every category from animal and artifact domains, a correlation value of the human ratings of typicality and the degrees of typicality computed by the formula typ_{sim} was observed, and this has been performed with the ten selected similarities sim. These correlation values serve as a performance indicator of how well a given typicality formula is able to predict the typicality ratings provided by human respondents. The well-known Kendall tau rank correlation coefficient τ_b was chosen to measure the this agreement, i.e., agreement between the list of category exemplars sorted by typicality. The τ_b is well suited since it measures the ordinal association between two quantities in the range from 1 (same ordering) to -1 (inverse, i.e., opposite ordering). To account for ties in typicality, we used the τ_b variant provided in Python library [26].

In this scenario, we explore various questions. Firstly, we are interested in whether typ_w provides a better prediction of typicality in comparison to the similarity-based scheme, i.e., to the ten formulas obtained from the similarity-based scheme by taking the ten selected measures sim mentioned above. Secondly, is there any observable influence of the particular domain from which the data are? Namely, we observed noticeable differences between the animal and the artifact domains in our previous experimental work on prediction of typicality [5]. Last but not least, we want to observe the influence of category and exemplar attributes on the ability to predict human ratings of typicality.

3.3. Results of analyses

Note first that according to a commonly accepted interpretation, the values of τ_b may be interpreted as follows: $\tau_b \geq 0.3$, $0.2 \leq \tau_b < 0.3$, $0.1 \leq \tau_b < 0.2$, and $0.0 \leq \tau_b < 0.1$ indicate strong, moderate, weak, and very weak correlation, respectively (analogously for negative values).

Let us start with the question of whether typ_w provides a strictly better prediction of the typicality. Figure 1 provides the mean values of τ_b correlations of all computed typicalities to the human typicality ratings across all categories in both domains.[7] The two lines represent two sets of attributes available in the Dutch data:

---

sim of similarity which we selected for our experimental evaluation, were omitted in this table because of lack of space.

[7]Note, that since all of the correlations were positive, using the mean as aggregation function is

The category-based attributes and exemplar-based attributes. The red dashed line highlights typicality rating typ_{HJ}, which is based on human degrees of similarity, HJ, and can be understood as a benchmark of typicality based on the similarity-based scheme (3).

The results provide strong experimental support for typ_w being superior in providing accurate predictions of typicality. The obtained mean correlation values are higher than those for any of the similarity-based typicalities, and, a fortiori, also for the original Rosch and Mervis' formula. This result seems remarkable because our previous extensive experimental research did not find a similarity measure that would provide a more accurate typicality prediction than the group of similarity coefficients Co1, RR, int, Di2, and CT3 [5]. All of these similarity-based predictions are weaker than the predictions based on typ_w typicality and confirm that information outside categories, as discussed in section 2.5, play an important role in the determination of typicality.

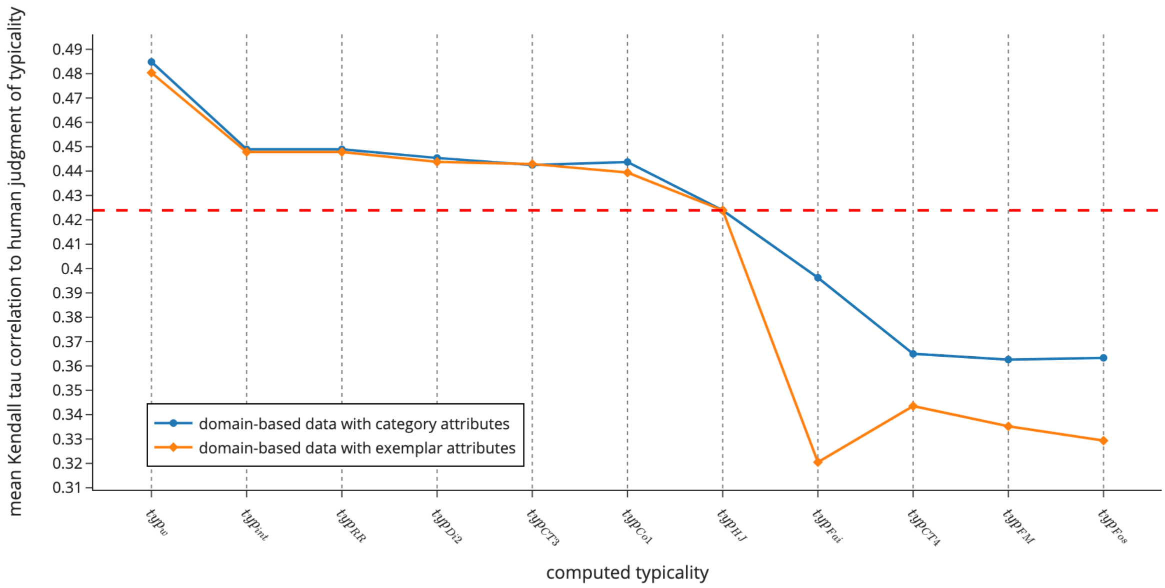

Figure 1: Mean correlations of computed typicality to the human judgment of typicality across all data.

The differences between the animal and artifact categories were observed and studied multiple times in the psychology of concepts [17]. To examine the influence

sound.

of a domain, we separated animal and artifact categories and plotted them in separate figures 2 and 3. These figures are organized in a similar fashion as figure 1 with the exception that correlations are calculated only on the animal and artifact domain data.

Our previous study found significant differences between animal and artifact domains. In the case of the animal domain, the typicality typ_{HJ} based on human similarity ratings HJ provides the strongest prediction of typicality [5]. As we can see in figure 2, the new typicality formula, typ_w, provides an even stronger prediction of typicality ratings than the previously best typicality typ_{HJ} based on human similarity ratings. The performance drop in the mean τ_b correlations between typ_{HJ} and the rest of the calculated degrees of typicality based on the similarity coefficients Co1, RR, int, Di2, and CT3 makes typ_w even more important. The significance of this observation consists in the fact that we previously argued that attributes provided by human respondents fail to fully describe the animal domain, and thus, the HJ similarity provides a better prediction of typicality. The performance on the artifact domain, presented in figure 3, follows a similar trend as in the case of typicality based on similarity measures. Nevertheless, the correlations are overall stronger than for the animal domain and the typ_w provides the strongest correlation from all presented typicality ratings.

Let us now examine the influence of attribute kinds, i.e., category vs. exemplar. This topic was addressed multiple times in the literature [17], and the usual consensus seems to be that the category and exemplar attributes provide different qualities. We previously found a support for this phenomenon in the case of typicality when computed using the similarity-based scheme [5]. In the case of domain data which describes the context of categories in a broader sense, the new typicality formula typ_w provides much more stable results, which are not influenced by the origin of the attributes. It seems worth noting that this was not always the case for the similarity-based typicalities.

4. Conclusions

In our paper, we reconsider Rosch and Mervis' definition of typicality of exemplars in human categories. We argue that the definition leads to two distinct schemes to compute degrees of typicality. While one was proposed in Rosch and Mervis' seminal paper, the other involves a measure of similarity and has not been expounded in the psychological literature. Our experimental evaluation utilizing the Dutch data demonstrates high correlation of both schemes exhibit high correlation with a human judgment of typicality. We examine their relationship and prove that Rosch and

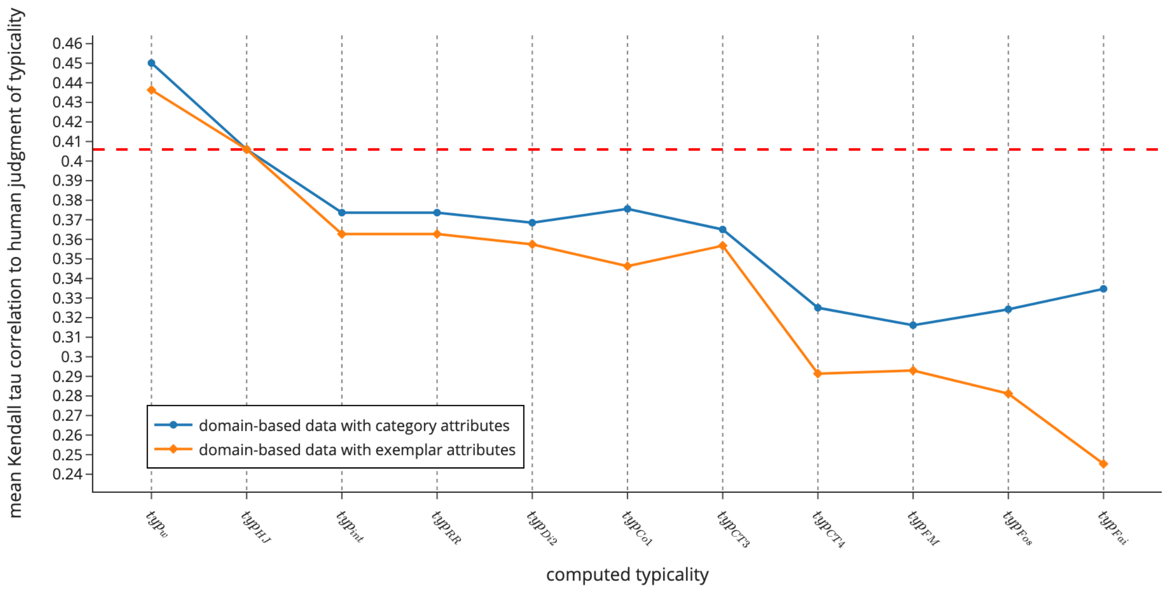

Figure 2: Mean correlations of computed typicality to the human judgment of typicality across categories of the animal domain.

Mervis' typicality formula, as well as its natural generalization, which takes into account the absence of attributes in a category rather than restricting to presence like the original formula, are equivalent to a particular case of the similarity-based formula with a properly chosen similarity measure. We then propose a technically simple but conceptually significant extension of Rosch and Mervis' formula that involves the concept of attribute characteristicness. We provide an extensive experimental evaluation of the considered typicality formulas and establish that the new typicality formula outperforms the original Rosch and Mervis' formula, the variants, as well as the similarity-based scheme. Our approach hence provides a novel psychological account of typicality, which is more effective than the previous ones and which is worth further psychological exploration.

References

[1] Ahn, W. K., Kim, N. S., Lassaline, M. E., Dennis, M. J.: Causal status as a determinant of feature centrality. Cognitive Psychology 41, no. 4, 361–416 (2000).

[2] Barsalou, L. W.: Ideals, central tendency, and frequency of instantiation as

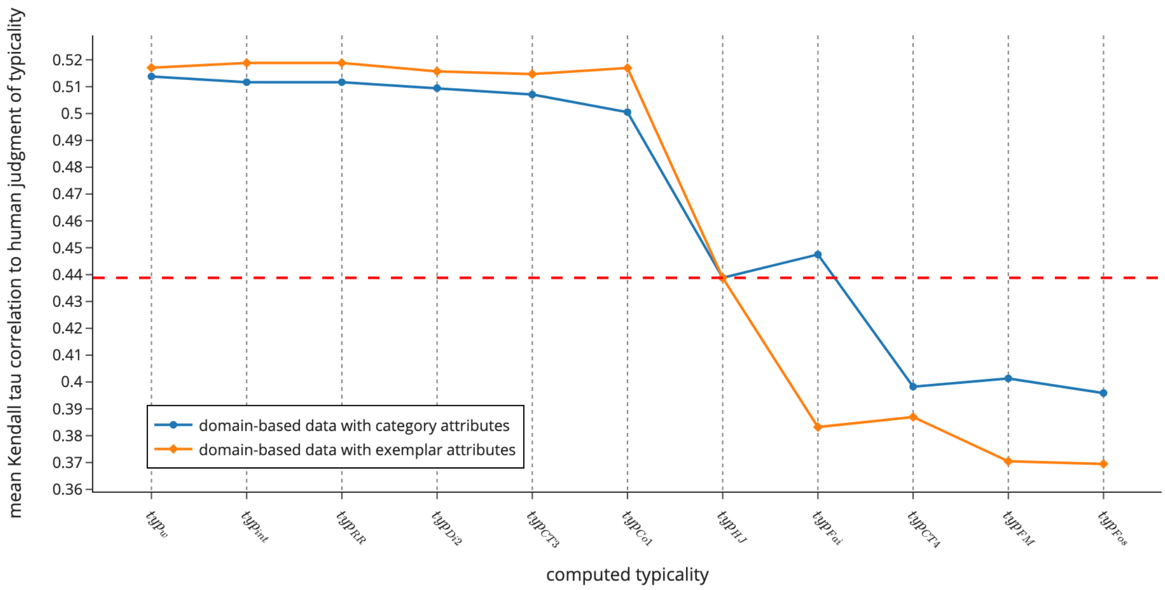

Figure 3: Mean correlations of computed typicality to the human judgment of typicality across categories of the artifact domain.

determinants of graded structure in categories. J. Experimental Psychology: Learning, Memory, and Cognition 11, no. 4, 629–654 (1985).

[3] Belohlavek, R., Mikula, T.: Typicality: A formal concept analysis account. International Journal of Approximate Reasoning 142, 349–369 (2022).

[4] Belohlavek, R., Mikula, T.: Dutch Concepts. GitHub repository (2023). `https://github.com/mikulatomas/dutch-concepts`.

[5] Belohlavek, R., Mikula, T.: Similarity metrics vs human judgment of similarity for binary data: Which is best to predict typicality? Applied Soft Computing 153, 111270 (2024).

[6] Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., Rogers, T. T., Hodges, J. R.: A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. Cognitive neuropsychology 20, no. 1, 27–47 (2003).

[7] Brusco, M., Cradit, J. D., Steinley, D.: A comparison of 71 binary similarity coefficients: the effect of base rates. PLOS ONE 16, no. 4, 1–19 (2021).

[8] Chin-Parker, S., Ross, B. H.: Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. Journal of Experimental Psychology: Learning, Memory, and Cognition 30, no. 1, 216 (2004).

[9] Choi, S. S., Cha, S. H., Tappert, C. C.: A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics 8, no. 1, 43–48 (2010).

[10] Dieciuc, M. A., Folstein, J. R.: Typicality: Stable structures and flexible functions. Psychonomic Bulletin & Review 26, 491–505 (2019).

[11] De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., Storms, G.: Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. Behavior Research Methods 40, no. 4, 1030–1048 (2008).

[12] Fisher, D. H.: A computational account of basic level and typicality effects. Proc. AAAI-88, 233–238 (1988).

[13] Garrard, P., Lambon Ralph, M. A., Hodges, J. R., Patterson, K.: Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. Cognitive neuropsychology 18, no. 2, 125–174 (2001).

[14] Hubálek, Z.: Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biological Reviews 57, no. 4, 669–689 (1982).

[15] Hsu, N. S., Schlichting, M. L., Thompson-Schill, S. L.: Feature diagnosticity affects representations of novel and familiar objects. Journal of Cognitive Neuroscience 26, no. 12, 2735–2749 (2014).

[16] Mervis, C. B., Catlin, J., Rosch, E.: Relationships among goodness-of-example, category norms, and word frequency. Bulletin of the Psychonomic Society 7, no. 3, 283–284 (1976).

[17] Murphy, G. L.: The Big Book of Concepts. MIT Press, Cambridge, Mass. (2002).

[18] Rosch, E.: Principles of categorization. In Rosch, E., Lloyd, B. B. (Eds.): Cognition and Categorization, Erlbaum, Hillsdale, NJ, 27–48 (1978).

[19] Rosch, E., Mervis, C. B.: Family resemblances: Studies in the internal structure of categories. Cognitive Psychology 7, no. 4, 573–605 (1975).

[20] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., Boyes-Braem, P.: Basic objects in natural categories. Cognitive Psychology 8, no. 3, 382–439 (1976).

[21] Russell, P. F., Rao, T. R.: On habitat and association of species of anopheline larvae in south-eastern Madras. Journal of the Malaria Institute of India 3, no. 1, 153–178 (1940).

[22] Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., Storms, G.: Dutch norm data for 13 semantic categories and 338 exemplars. Behavior Research Methods, Instruments, & Computers 36, no. 3, 506–515 (2004).

[23] Sokal, R. R., Michener, C. D.: A statistical method for evaluating systematic relationships. The University of Kansas Science Bulletin 38, part 2, no. 22, 1409–1438 (1958).

[24] Sloman, S. A., Love, B. C., Ahn, W. K.: Feature centrality and conceptual coherence. Cognitive Science 22, no. 2, 189–228 (1998).

[25] Sloman, S. A., Ahn, W. K.: Feature centrality: Naming versus imagining. Memory & Cognition 27, no. 3, 526–537 (1999).

[26] Virtanen, P., Gommers, R., Oliphant, T. E. et al.: SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17, no. 3, 261–272 (2020).

[27] Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., Willett, P.: Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. Journal of Chemical Information and Modeling 52, no. 11, 2884–2901 (2012).

[28] Ward, T. B., Dodds, R. A., Saunders, K. N., Sifonis, C. M.: Attribute centrality and imaginative thought. Memory & Cognition 28, 1387–1397 (2000).

[29] Warrens, M. J.: Similarity Coefficients for Binary Data. PhD thesis, Leiden University, The Netherlands (2008). `https://scholarlypublications.universiteitleiden.nl/access/item%3A2961378/view`.

[30] Wijaya, S. H., Afendi, F. M., Batubara, I., Darusman, L. K., Altaf-Ul-Amin, M., Kanaya, S.: Finding an appropriate equation to measure similarity between binary vectors: Case studies on Indonesian and Japanese herbal medicines. BMC Bioinformatics 17, 520 (2016).

[31] Woollams, A. M.: Apples are not the only fruit: The effects of concept typicality on semantic representation in the anterior temporal lobe. Frontiers in Human Neuroscience 6, 85 (2012).

[32] Yeh, W., Barsalou, L. W.: The situated nature of concepts. The American Journal of Psychology 119, no. 3, 349–384 (2006).

Appendix: Proof of theorem 1

In the proof, we denote by val(φ) the truth value of proposition φ. That is, val(xIy) = 1 if x has y, and val(xIy) = 0 if x does not have y. In addition, we use the notation introduced in section 2, i.e., $\{x\}\uparrow$ is the set of all attributes possessed by x, and $|S|$ is the number of elements of the set S.

We have

$$\begin{aligned} & |A|\cdot|Y|\cdot typ_{\mathrm{SMC}^{a^+,a^-}}(x,A) = |A|\cdot|Y|\cdot\frac{\sum_{x_1\in A} sim^{a^+,a^-}_{\mathrm{SMC}}(x,x_1)}{|A|} \\ = \; & |A|\cdot|Y|\cdot\frac{\sum_{x_1\in A}\frac{a^+\cdot|\{x\}^\uparrow\cap\{x_1\}^\uparrow|+a^-\cdot|Y-(\{x\}^\uparrow\cup\{x_1\}^\uparrow)|}{|Y|}}{|A|} \\ = \; & a^+\cdot\sum_{x_1\in A}|\{x\}^\uparrow\cap\{x_1\}^\uparrow| + a^-\cdot\sum_{x_1\in A}|Y-(\{x\}^\uparrow\cup\{x_1\}^\uparrow)|, \end{aligned} \tag{10}$$

For the first term in (10) we obtain

$$\begin{aligned} & a^+\cdot\sum_{x_1\in A}|\{x\}^\uparrow\cap\{x_1\}^\uparrow| = a^+\cdot\sum_{x_1\in A}|\{y\in Y\,;\, xIy \text{ and } x_1Iy\}| \\ = \; & a^+\cdot\sum_{x_1\in A}\sum_{y\in Y:xIy}\mathrm{val}(x_1Iy) = a^+\cdot\sum_{y\in Y:xIy}\sum_{x_1\in A}\mathrm{val}(x_1Iy) \\ = \; & a^+\cdot\sum_{y\in Y:xIy}|\{x_1\in A\,;\, x_1Iy\}| = a^+\cdot\sum_{y\in Y:xIy} w^+(y,A). \end{aligned}$$

For the second term in (10), we have

$$
\begin{aligned}
& a^- \cdot \sum_{x_1 \in A} |Y - (\{x\}^\uparrow \cup \{x_1\}^\uparrow)| = a^- \cdot \sum_{x_1 \in A} |Y \cap (\overline{\{x\}^\uparrow} \cap \overline{\{x_1\}^\uparrow})| \\
= \; & a^- \cdot \sum_{x_1 \in A} |\{y \in Y\,;\, x\overline{I}y \text{ and } x_1\overline{I}y\}| \\
= \; & a^- \cdot \sum_{x_1 \in A} \sum_{y \in Y : x\overline{I}y} \mathrm{val}(x_1\overline{I}y) = a^- \cdot \sum_{y \in Y : x\overline{I}y} \sum_{x_1 \in A} \mathrm{val}(x_1\overline{I}y) \\
= \; & a^- \cdot \sum_{y \in Y : x\overline{I}y} |\{x_1 \in A\,;\, x_1\overline{I}y\}| = a^- \cdot \sum_{y \in Y : x\overline{I}y} w^-(y, A).
\end{aligned}
$$

The reasoning above thus yields

$$
\begin{aligned}
& |A| \cdot |Y| \cdot \mathit{typ}_{\mathrm{SMC}^{a^+,a^-}}(x, A) \\
= \; & a^+ \cdot \sum_{x_1 \in A} |\{x\}^\uparrow \cap \{x_1\}^\uparrow| + a^- \cdot \sum_{x_1 \in A} |Y - (\{x\}^\uparrow \cup \{x_1\}^\uparrow)| \\
= \; & a^+ \cdot \sum_{y \in Y : xIy} w^+(y, A) + a^- \cdot \sum_{y \in Y : x\overline{I}y} w^-(y, A) = \mathit{typ}^{a^+,a^-}_{\mathrm{RM}^\pm}(x, A),
\end{aligned}
$$

proving the theorem. $\square$

typ_w-order	typ_w	typ_{RR}-order	typ_{RR}	typ_{HJ}-order	typ_{HJ}
sparrow	29.3104	parrot	0.0921	dove	0.4944
robin	29.1250	sparrow	0.0906	magpie	0.4899
chickadee	29.0273	dove	0.0900	swallow	0.4834
blackbird	28.9722	chickadee	0.0893	woodpecker	0.4766
parrot	28.9669	blackbird	0.0888	cuckoo	0.4750
parakeet	28.7782	crow	0.0884	chicken	0.4738
dove	28.5193	cuckoo	0.0880	seagull	0.4735
crow	28.2343	robin	0.0880	crow	0.4730
canary	28.1999	parakeet	0.0872	blackbird	0.4714
magpie	27.7615	owl	0.0870	falcon	0.4700
cuckoo	27.6748	falcon	0.0868	heron	0.4691
swallow	27.6534	duck	0.0862	pheasant	0.4683
falcon	27.3447	magpie	0.0859	sparrow	0.4679
woodpecker	27.2694	swallow	0.0851	turkey	0.4677
owl	26.8699	woodpecker	0.0846	parakeet	0.4601
chicken	26.3930	canary	0.0838	chickadee	0.4592
eagle	26.3503	seagull	0.0838	duck	0.4554
rooster	26.2568	eagle	0.0834	robin	0.4541
seagull	26.2083	rooster	0.0825	parrot	0.4538
duck	25.8912	chicken	0.0824	rooster	0.4511
turkey	25.6492	turkey	0.0810	peacock	0.4510
pheasant	25.6242	pheasant	0.0800	canary	0.4507
stork	25.3355	vulture	0.0796	owl	0.4427
heron	24.8637	stork	0.0795	vulture	0.4387
vulture	24.0218	heron	0.0788	pelican	0.4378
peacock	23.9239	ostrich	0.0777	stork	0.4368
swan	22.5596	swan	0.0755	swan	0.4356
ostrich	22.1024	peacock	0.0755	eagle	0.4339
pelican	21.9538	pelican	0.0739	ostrich	0.4025
penguin	20.6356	penguin	0.0716	penguin	0.3286

Table 3: Typicality ratings for “bird” for the animal domain with exemplar attributes.

Appendix D

Comparing similarity measures for binary data with human judgment of similarity

In this preprint (Belohlavek and Mikula, 2024d), we carefully gathered and documented 69 similarity measures, which were directly compared to the human judgment of similarity provided in the Dutch data. The results of these experiments are briefly described in Section 4.5. The preprint results from joint research with my supervisor Radim Bělohlávek.

Comparing similarity measures for binary data with human judgment of similarity

Radim Belohlavek, Tomas Mikula

Department of Computer Science, Palacký University Olomouc, Czech Republic

Abstract

Since the late nineteenth century, more than seventy similarity measures for binary data have been proposed. Several studies examined mutual relationships between the existing measures. Unlike those studies, we explore a different question. We investigate to what extent the existing similarity measures agree with human judgment of similarity. We utilize now available psychological data involving objects from two large semantic domains, their description by binary attributes, and human ratings of similarity of these objects. We consider sixty-nine similarity measures, which we collected in the literature, and study using rank-order correlation the agreement of these similarity measures with human ratings of similarity. Our most important finding is that while most of the similarity measures exhibit a reasonable correlation with human similarity, the correlation strength delineates several groups of measures that consistently display a similar relationship to human similarity across the examined domains. We analyze common properties of the measures in the revealed groups, discuss factors affecting the correlation strength, and compare the groups with classifications of similarity measures observed in the literature.

Keywords: similarity, human judgment, similarity measure, binary data, psychology

1. Our aim

In many fields, the objects of interest are described by binary (yes-no, presence-absence) attributes, i.e., features that any given object either does or does not possess. Collections of objects and their descriptions by binary attributes constitute

Email addresses: radim.belohlavek@acm.org (Radim Belohlavek), mail@tomasmikula.cz (Tomas Mikula)

what is referred to as binary data. For many tasks, such as classification, categorization, or clustering, measuring the similarity of objects in binary data is crucial. Since the late 19th century, over seventy measures of similarity for binary data have been proposed for applications in areas as diverse as ecology, ethnology, biology, psychology, taxonomy, pattern recognition, and data mining.

Both formal and empirical properties of the proposed measures have been investigated in numerous publications. Several of them explored mutual relationships between individual similarity measures; see, e.g., [7, 8, 24, 29, 55, 59, 58] for some influential and recent studies. Those studies typically involve tens of measures, and a comparison of these measures is performed on data from a particular domain of interest, such as chemistry or ecology, as well as on random data.

Unlike those studies, we pursue a different question. Namely, we explore how the existing similarity measures agree with human judgment of similarity. This question is critical because the psychological plausibility of similarity measures is an obvious imperative. Yet, the question has not been seriously examined before, primarily because appropriate psychological data was unavailable.

In our study, we utilize now available high-quality psychological data [12], which involves categories (concepts) and exemplars (objects) from two large semantic domains. The data also provides human ratings of similarity of these exemplars. In addition, it provides descriptions of these objects by carefully solicited binary attributes. Using these attributes, degrees of similarity of the exemplars may be computed for any given measure of similarity. We consider 69 similarity measures we collected in the literature and present them in detail, along with various remarks on the flaws we found. To our knowledge, this is the most extensive collection examined in the literature and may be utilized in further studies. Our assessment of agreement of a similarity measure with a human rating of similarity is conducted in an ordinal fashion: Rather than comparing directly the degrees of similarity, i.e., those computed by a similarity measure and those provided by human judgment, we employ rank-order correlation. Hence, we consider the extent to which the calculated similarity agrees with human similarity on whether any given pair of objects is more similar than another.

Our paper is organized as follows. In section 2, we present preliminaries on similarity measures along with the 69 measures used in our study. The data we use, the rationale of our experiments, and the experimental results are presented in section 3. Section 4 provides conclusions and outlines future research topics.

2. Similarity measures

2.1. The concept of similarity measure for binary data

According to a common understanding, which we use to cover a variety of particular instances, a measure of similarity of objects in a set X is a binary function

$$sim : X \times X \to \mathbb{R}$$

with the values $sim(x, y)$ interpreted as the extent (degree) to which x is similar to y.

Remark 1. Additional constraints, such as symmetry, i.e., $sim(x, y) = sim(y, x)$, maximality, i.e., $sim(x, y) \leq sim(x, x)$, or various dual forms of the triangle inequality, are often considered as they are satisfied by several similarity measures. We do not impose them to cover the wide variety of similarity functions proposed in the literature.

Restrictions on the range of sim may also be imposed, such as $sim(x, y) \in [0, 1]$ or $sim(x, y) \in [-1, 1]$, which may be circumvented by appropriate scaling functions. We allow for general functions $sim : X \times X \to \mathbb{R}$ for simplicity. This does not present a problem in our treatment because we only consider ordinal information provided by the similarity measures. In particular, our correlation analysis only takes into account whether $sim(x, y) \leq sim(x', y')$ or not for the considered pairs of objects $\langle x, y\rangle$ and $\langle x', y'\rangle$, rather than the actual values $sim(x, y)$ and $sim(x', y')$.

For convenience, the set X of all objects described by n binary attributes is commonly identified with the set $\{0, 1\}^n$ of all n-dimensional binary vectors. For example, the vector

$$x = \langle 1, 0, 0, 1, 1\rangle$$

in $x \in \{0, 1\}^5$ represents the object described by 5 binary attributes for which $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, $x_4 = 1$, and $x_5 = 1$. In other words, the object has the first, fourth, and fifth attribute but not the second nor the third.

Let now $x, y \in \{0, 1\}^n$ be two binary vectors. Considerations of similarity conveniently utilize the following scheme:

	$y = 1$	$y = 0$	Σ
$x = 1$	a	b	$a + b$
$x = 0$	c	d	$c + d$
Σ	$a + c$	$b + d$	n

(1)

in which, e.g., a is the number of attributes i for which $x_i = 1$ and $y_i = 1$, b is the number of i for which $x_i = 1$ and $y_i = 0$, $a + c$ is the number of attributes for which $y_i = 0$, etc. Clearly, $n = a + b + c + d$.

For example, for the vectors

$$x = \langle 0,1,1,0,0,1,0,1,1,0\rangle \quad \text{and} \quad y = \langle 1,1,1,0,0,1,0,1,0,1\rangle$$

in $\{0,1\}^{10}$, we obtain

	$y=1$	$y=0$	Σ
$x=1$	4	1	5
$x=0$	2	3	5
Σ	6	4	10

The similarity measures considered in the literature are naturally defined by a formula involving a, b, c, and d, which correspond to $x, y \in \{0,1\}^n$ according to (1). For example, the well-known simple matching coefficient (SMC) and the Jaccard measure (Jac) are defined by

$$sim(x,y) = \frac{a+d}{a+b+c+d} \quad \text{and} \quad sim(x,y) = \frac{a}{a+b+c}, \tag{2}$$

respectively.

2.2. Similarity measures employed in our study

We consider 69 widely known similarity measures that appear in the literature, e.g., in the comparative studies [7, 8, 29, 55, 59]. The employed measures are described in table 5 in the appendix. For each measure, the table contains its abbreviation, its name, a formula for computing the values $sim(x,y)$ from the corresponding numbers a, b, c, and d, and a reference to the work in which the measure was introduced.[1] The measures are sorted alphabetically that one can quickly find the measures when analyzing our experimental results.

Remark 2. Some of the measures, namely, Ku1, Mou, Pe3, Pr3, SS5, Sti, and Tar (see table 5) are not defined for certain combinations of a, b, c, and d that occur with our data. These measures are omitted from the results presented in section 3.

Notice that some of the involved similarity measures are undefined even for a pair of objects having the same attributes, i.e. yielding $b = c = 0$. This counterintuitive property can be remedied by redefining each such measure to attain its maximal value for $b = c = 0$. We, however, avoid this option to stick to the definitions presented in the literature.

[1] As regards references to the original works, we performed a comprehensive search of the literature and checked all the referenced works, rather than adopting the sometimes unreliable information provided by the available comparative studies. In cases we were not able to identify the original paper, the table contains reference(s) to the comparative studies that include the measure, preceded by $*$.

2.3. Shortcomings found in the literature

We found several inaccuracies and flaws in the comparative papers on similarity measures for binary data [7, 8, 29, 55, 59]. The following list presents the significant ones.

1. AC: [29] lists a slightly different formula for AC, namely $\frac{1}{50\pi}\sqrt{\frac{a+d}{n}}$, i.e., a formula yielding a value 100× smaller than our formula. The original paper [3] presents the formula we employ.
2. Col: [8], [29], [59] list a different formula, namely $\frac{\sqrt{2}(ad-bc)}{\sqrt{(ad-bc)^2-(a+b)(a+c)(b+d)(c+d)}}$. This formula also appears in [10] on p. 416 as a so-called mean square contingency, but is not meant as the similarity measure which the authors present in their paper.
   The Abydos library [1] lists our formula for Col as the Cole similarity measure, but the conditions corresponding to the three parts of the formula, which are listed in [1] are not mutually exclusive, hence the formula in the Abydos library is ambiguous. Note also that Co2 appears in [10] on p. 420 as C_7, but is only meant in [10] as a part of the formula we denote Col; Co1 is symmetric to Co2 but does not appear in [10].
3. Eyr: [29] lists a different formula, namely $\frac{a-(a+b)(a+c)}{(a+b)(a+c)(b+d)(c+d)}$.
   In the original paper [15], however, we have not found any of the two formulas.
4. FM: [7], [8], [29], and [59] list a different formula, $\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{\max(a+b,a+c)}{2}$. This is apparently a wrong formula. Namely, the original paper [18] contains the formula we use as FM, and notes that this formula results as a modification of a formula used in [17].
5. Fos: [7] lists a different formula, $\frac{n(a-\frac{1}{2})^2}{\sqrt{(a+b)(a+c)}}$. This is a misprint as the paper to which the authors refer contains a different formula, namely the one we use. Note also that the paper referred to by [7] is: Holliday, J. D., Chu, C.-Y., Willett, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit strings. *Combin. Chem. High-Throughput Screening* 2012, 5, 155–166. The remaining comparative works which employ the Fossum measure, [8], [55], and [59], all refer to the above paper by Holliday et al., while a reference to the original paper is missing in [8], [55], and [59], as well as in the paper by Holliday. In fact, the Fossum measure is proposed in [21, p. 65, formula (7)].
6. GW: [7], [8], [29], and [59] list a different formula, namely $\log a - \log n - \log(\frac{a+b}{n}) - \log(\frac{a+c}{n})$; [7] and [29] refer to [22], [8] does not contain a reference for

this measure, and refers to [8] and [29]. In fact, [22] includes our formula, as does [1]. Note also that [1] mentions that most comparative papers include the different formula, which we mentioned above, and that "neither this formula nor anything similar or equivalent to it appears anywhere within the cited work" of Gilbert and Wells [22]. This, however, is not the case.

Namely, [22] states that if all species in their model are distributed independently of each other, one has $n_{ij} = Nq_iq_j$, where—in our notation—$n_{ij} = a$, $N = n$, $q_i = \frac{a+b}{n}$, and $q_j = \frac{a+c}{n}$. Note that q_i and q_j are not explicitly described in [22], but it follows from the considerations of the authors that q_i and q_j are the probability of having some attribute by object i and j, respectively, i.e. $q_i = \frac{a+b}{n}$ and $q_j = \frac{a+c}{n}$. Then, clearly, $\log a = \log n + \log \frac{a+b}{n} + \log \frac{a+c}{n}$, hence $\log a - \log n - \log \frac{a+b}{n} - \log \frac{a+c}{n}$ may be understood as a deviation from 0.

7. Gle: [59] lists an additional measure called NeiLi. The authors did not realize that their formula for NeiLi yields the same value as Gle, which they list separately.
8. Jac: Both [8] and [59] list separately the so-called Tanimoto measure with the formula $\frac{a}{(a+b)+(a+c)-a}$. This formula clearly yields the same value as Jac.
9. Ku2: The authors in [8] have not realized that Ku2 yields the same value as the formula $\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$, which they list separately as the Driver-Kroeber measure. In [29], the author lists measures A7 and A8, which he calls Kulczynski and Drive & Kroeber measures, respectively; the formula for A7 coincides with our formula for Ku2; the one for A8 yields the same value as A7.

   Note also that [7], [8], [29], [55], and [59] list the so-called Johnson measure with a formula $\frac{a}{a+b} + \frac{a}{a+c}$. Clearly, this formula yields the value of $2 \cdot \mathrm{Ku2}$, hence we do not include the Johnson measure [31].
10. SS2: All [7], [8], [29], [55], and [59] list separately the so-called Gower-Legendre measure with the formula $\frac{a+d}{a+0.5(b+c)+d}$. This formula, however, yields the same value as SS2, which is not mentioned in these works. We hence omit the Gower-Legendre measure in our list.
11. SS3: [8] contains a misprint in the formula for SS3.
12. SS4: Both [8] and [59] include this measure as Ochiai 2 and list separately, apparently with a misprint, an equivalent formula called the Sokal-Sneath 5 measure.
13. Sti: [7], [8], and [59] list a different formula, namely $\log \frac{n(|ad-bc|-n/2)^2}{(a+b)(a+c)(b+d)(c+d)}$. This formula is different from the one which we use and which comes from the original paper [53]. The Abydos library [1] uses our formula.
14. Tar: Both [8] and [59] separately list another measure called the Ample measure

with a formula which yields— given that a, b, c, and d are nonnegative—the same value as Tar.

3. Experiments and results

3.1. Data

The Dutch data [12] is the most comprehensive data concerning human categories and accompanying phenomena. The availability of quality data obtained from humans is vital for experiments to be of psychological significance. The Dutch data is unique in this regard due to its considerable scope. It is substantially more extensive than the previously available psychological data of a similar kind. The data has been collected by psychologists of the Katholieke Universiteit Leuven as part of a study involving hundreds of human respondents, and builds on a previous study conducted at the KU Leuven [46]. It makes available information regarding common language categories, binary features (yes/no attributes, binary variables) relevant to these categories, exemplars (objects) in these categories, information on similarity of these exemplars, as well as further characteristics.

The Dutch data involves sixteen linguistic categories. Each category is represented by several exemplars, such as a deer for the category "mammal." The exemplars were obtained in an exemplar-generating process involving 527 participants. For each category, a number—aimed at 30 per category—of exemplars was selected, including typical and atypical ones, with the restriction that they be familiar to the vast majority of participants [46]. The data includes ten natural kinds and six artifact categories. The natural kind categories are:[2] "fruit" (with 30 exemplars); "vegetables" (30); "professions" (30); "sports" (30); the animal categories "amphibians"(5),[3] "birds" (30), "fish" (23), "insects" (26), "mammals" (30), and "reptiles" (22).[4] The artifact categories are: "clothing" (29), "kitchen utensils" (33), "musical instruments" (27), "tools" (30), "vehicles" (30), and "weapons" (20). Altogether, the categories comprise 249 exemplars for the natural kind and 166 exemplars for

[2]In this paragraph, we use the plural in category names, as the authors do [12]; below, we use the singular, i.e. "bird" rather than "birds" for consistency with our previous writings.

[3]This category contains just 5 exemplars. As these exemplars are included in the category "reptiles", we omit it in our considerations below. The reasons to include the exemplars of "amphibians" in "reptiles" are explained in [12].

[4]The exemplar-by-feature applicability matrices, which we describe below and use in our experiments, contain only 20 exemplars of the category "reptiles" because the respondents who were to fill in these matrices turned out to not to be familiar with two exemplars, iguanodon and komodo. We hence exclude these two exemplars from our experiments.

the artifact categories, which are representative of the respective categories.[5] The categories have considerable coverage of the respective domains; for example, the animal categories cover a large part of the known animal domain.

An important part of the Dutch data is represented by similarity data. In particular, pairwise similarity ratings were collected in a process involving 92 respondents, who were asked to provide similarity ratings for pairs of exemplars of all the included categories except for "amphibian," whose five exemplars are included in "reptiles." For ten of these categories, similarity judgments were available and hence adopted from the previous study [46] involving 42 participants, but new judgments were obtained even for these ten categories to ensure the desired reliability. For each category, the data thus contains reliable similarity ratings for each pair of category exemplars (the cross-category similarity ratings are not considered). The ratings are on the scale 1 (totally dissimilar) to 20 (totally similar).

Another part of the data vital for our purpose consists of descriptions of the involved exemplars by features, i.e., binary attributes. These descriptions are represented by the so-called exemplar-by-feature applicability matrices in the Dutch data, which we describe in the next paragraph. The attributes were obtained by a process described in [12]. In particular, the attributes were generated by $1,003$ respondents in two ways: Respondents were asked to list relevant attributes for a given category, which are called the category attributes. In addition, other respondents were asked to list relevant attributes for each object (i.e., exemplar) involved in the data, resulting in what is called the exemplar attributes. Unions of all the exemplar attributes listed for all the objects in a given category are then considered, as well as the union of all exemplar attributes of all the objects in the animal domain and the union of exemplar attributes for the artifact domain.

The exemplar-by-feature applicability matrices are various kinds of binary matrices in which the rows and the columns correspond to some of the objects and the attributes, respectively, and the entries contain information about whether a particular object has or does not have a particular attribute. Altogether 77 respondents filled these matrices. Each matrix was completed by 4 different respondents. The data also contains the corresponding aggregated matrices, in which the values, viz. 0, 1, 2, 3, and 4, indicate the number of respondents who agreed that the respective object has the respective attribute. To obtain binary matrices, and thus data described by binary attributes, from these aggregated matrices, one naturally thresholds the

[5] In addition to the five amphibians included in reptiles and two omitted exemplars of reptiles (see above), three exemplars of artifact categories are included in two distinct categories.

matrix entries. We present our experiments for the threshold equal to 2.[6] Hence, our binary matrices contain 1 in the entry corresponding to the object x and the attribute y if at least two respondents agreed that x has y.

In particular, we use the binary matrices described in table 1 and table 2. For example, the first row in table 1 refers to two binary matrices. The first one is a 30×28 matrix describing which of the 30 exemplars of the category "bird" have which of the 28 category attributes of this category, i.e., attributes listed as category attributes for "bird" by the respondents. The second is a 30×225 matrix telling which of the 30 exemplars of "bird" have which of the 225 exemplar attributes for this category, i.e., the attributes listed as exemplar attributes for some exemplar of "bird". In the same vein, the 166×301 binary matrix referred to by the second row in table 2 describes which of the 166 exemplars in the artifact domain have which of the corresponding 301 category attributes; the 166 objects are all the objects of the categories "clothing", "kitchen utensil", "musical instrument", "tool", "vehicle", and "weapon", and the 301 category attributes are all attributes listed as category attributes for these six categories. Likewise, the $166 \times 1,295$ matrix describes which of the exemplars in the artifact domain have which of the corresponding $1,295$ exemplar attributes, i.e., all the attributes listed as exemplar attributes for some of the 166 exemplars in the artifact domain.

To facilitate ease of machine processing, we adjusted the data to remove some minor semantic and technical faults and inconveniences. Our adjustment includes fixing misspelled object and attribute names, correcting errors in the English translation, removing duplicated English names, and converting all object and attribute names to lowercase. We also fixed the invalid CSV format of some files. The corrected version of Dutch data is easily machine-processable and is publicly available on GitHub [5] along with a convenient Python wrapper.

3.2. Experiments

3.2.1. Comparing similarity measures with human judgment

Our experiments aim to explore how well the various similarity measures agree with a human judgment of similarity. As explained in the previous section, the Dutch data provides, for each category, human similarity ratings $\mathrm{HJ}(x, y)$ for each pair x and y of exemplars in this category. Since the Dutch data contains descriptions

[6] While thresholds 1 (at least one respondent agrees) and 4 (all respondents agree) may arguably be considered extreme, we regard both thresholds 2 and 3 as reasonable choices. The results for the threshold equal to 3 are similar to those we obtained for the threshold 2, and we do not present them due to lack of space.

category	objects	category attributes	exemplar attributes
bird	30	28	225
clothing	29	38	258
fruit	30	32	233
fish	23	32	156
insect	26	37	214
kitchen utensil	33	39	328
mammal	30	34	288
musical instrument	27	39	218
profession	30	21	370
reptile	20	35	179
sport	30	33	382
tool	30	37	285
vegetable	30	30	291
vehicle	30	34	322
weapon	20	32	181

Table 1: Category-based binary matrices used in our experiments.

domain	objects	category attributes	exemplar attributes
animal	129	225	764
artifact	166	301	1,295

Table 2: Domain-based binary matrices used in our experiments.

of these exemplars by binary attributes, one may also compute the corresponding degrees $sim(x, y)$ for every considered similarity measure sim.

In fact, the Dutch data provides four collections of binary attributes that describe the exemplars of any given category, namely the collections of category attributes and exemplar attributes of the category-based and domain-based matrices referred to by tables 1 and 2, respectively. The similarity degrees $sim(x, y)$ may hence be computed with respect to these four attribute collections and the corresponding binary matrices.[7]

Let now sim be an arbitrary similarity measure. Consider a category, such as "mammal" represented by 30 exemplars, and some collection of binary attributes, such as the 288 exemplar attributes of this category. Using the 30×288 category-based matrix corresponding to "mammal," we compute the similarity degrees $sim(x, y)$ for all exemplars x and y of this category, and may hence compare the similarity measure sim with human similarity HJ. For each category, such a comparison of sim with HJ may be performed with respect to any of the four available collections of attributes.[8]

We approach the comparison of sim with HJ in an ordinal manner by asking to what extent sim and HJ agree on being more similar as regards the similarity of exemplar pairs. Such logic is conveniently implemented by the Kendall τ rank-order coefficient, which we apply to the list of all pairs $\langle x_i, x_j \rangle$ of exemplars ordered by the computed ratings $sim(x_i, x_j)$ and the list ordered by the human ratings $\mathrm{HJ}(x_i, x_j)$.[9] Call a couple consisting of two pairs of exemplars, $\langle x_i, x_j \rangle$ and $\langle x_k, x_l \rangle$, concordant if

$$\begin{array}{lll} sim(x_i, x_j) < sim(x_k, x_l) & \text{and} \quad \mathrm{HJ}(x_i, x_j) < \mathrm{HJ}(x_k, x_l), & \text{or} \\ sim(x_i, x_j) > sim(x_k, x_l) & \text{and} \quad \mathrm{HJ}(x_i, x_j) > \mathrm{HJ}(x_k, x_l), & \end{array}$$

i.e., if sim and HJ agree on being more similar for these pairs. If sim and HJ disagree, the couple is called discordant. The basic Kendall τ coefficient is defined as the ratio

$$\frac{\text{no. concordant couples} - \text{no. discordant couples}}{\text{no. all couples}},$$

[7]For instance, the exemplars of "mammal" are described by four collections of attributes: the 34 category attributes in the 30×34 category-based matrix, the 228 exemplar attributes in the 30×288 category-based matrix, the 225 category attributes in the 129×225 domain-based matrix, and the 764 exemplar attributes in the 129×764 domain-based matrix.

[8]Such as the above-mentioned 30-, 288-, 225-, and 764-attribute collections for the category "mammal."

[9]If the considered similarity measures were symmetric, i.e. verified $sim(x_i, x_j) = sim(x_j, x_i)$, it would be sufficient to consider only the pairs $\langle x_i, x_j \rangle$ with $i \leq j$. Since the similarity measures proposed in the literature include non-symmetric measures, we consider all pairs $\langle x_i, x_j \rangle$.

which ranges from -1 to 1. If $\tau = -1$, then all couples are discordant, i.e., the two lists of pairs of exemplars ordered by *sim* and HJ, respectively, are mutually inverse, i.e., represent opposite orderings. If $\tau = 1$, then all couples are concordant, i.e., the two lists are identical. To properly account for ties, i.e., cases in which $sim(x_i, x_j) = sim(x_k, x_l)$ or $\mathrm{HJ}(x_i, x_j) = \mathrm{HJ}(x_k, x_l)$, we employ the τ_b variant of the Kendall τ coefficient. We utilize the implementation of τ_b in a Python library [57].

In addition to representing a well-founded robust statistic, which also handles ties well, the Kendall coefficient provides—because of its straightforward consideration of concordant and discordant couples—an intuitively clear measure of ordinal agreement. This is why it is commonly regarded as a preferred rank-order correlation coefficient. We, nevertheless, also briefly report on the results obtained using the Spearman rank-order correlation, which is another well-known but less robust rank-order correlation.

3.2.2. Results

We now present the results of the experiments, whose scenario is described in section 3.2.1, using graphs depicting correlations between the examined similarity measures and the human judgment of similarity. The graphs are provided in figs. 1, 2, and 3 in this paper, and in figs. 6–13 in the online supplementary material.

The correlations are computed using the category-based binary matrices of table 1 that correspond to the 9 natural categories are displayed in figs. 6 (category attributes) and 7 (exemplar attributes); those corresponding to the 6 artifact categories are shown in figs. 8 (category attributes) and 9 (exemplar attributes). Analogously, the correlations based on the two domain-based matrices of table 2 that correspond to the categories of the animal domain are displayed in figs. 10 (category attributes) and 11 (exemplar attributes); those based on the two domain-based matrices of table 2 corresponding to the categories of the artifact domain are presented in figs. 12 (category attributes) and 13 (exemplar attributes). Figs. 1, 2, and 3 provide summarized views of these correlations.

In these graphs, the horizontal axis represents the examined similarity measures which are referred to by their abbreviations introduced in table 5. In each graph, the measures are sorted according to the average correlation values presented in the particular graph from the highest to the lowest. A "(d)" is appended to the abbreviation, as in "Fos (d)," if the absence of both attributes, which is represented by the number d in the contingency table (1), increases the value of the particular similarity measure. The vertical axis represents the values of Kendall τ correlation.[10]

[10] For instance, the red crosses in fig. 6 represent the correlation values of the particular similarity

The summarizing graphs in figs. 1, 2, and 3 are organized analogously.[11] For instance, the graph in fig. 1 displays the averages of correlations over all the categories of the animal domain using the individual category-based matrices as well as the animal-domain matrix for both the category and the exemplar attributes. That is, the graph represents a summary of the graphs in figs. 6, 7, 10, and 11 (the natural categories outside the animal domain are disregarded in this summary). Fig. 2 provides the same kind of information for the artifact domain and fig. 3 represents the averages over all the categories of both the animal and the artifact domain.

measures with the human judgment of similarity computed for all pairs of the 30 exemplars of the category "mammal" using the 30×34 binary matrix of table 1 describing these 30 exemplars by the 34 category attributes associated with "mammal."

[11] The horizontal axes of the summarizing graphs in figs. 1, 2, and 3 also depict a partitioning of the similarity measures into groups A–E that is addressed in section 3.2.3.

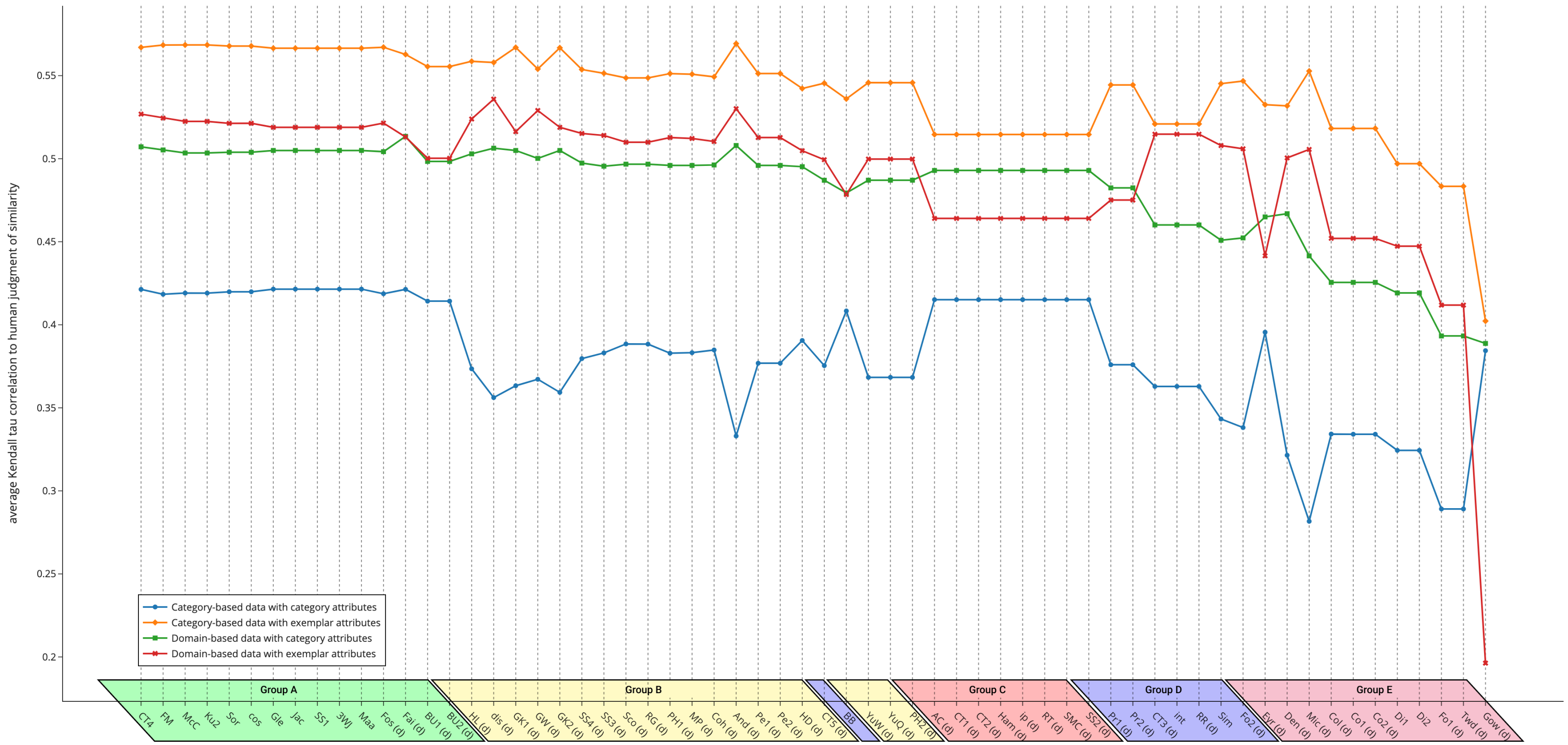

Figure 1: Average correlations of computed similarity to human judgment of similarity across categories of the animal domain.

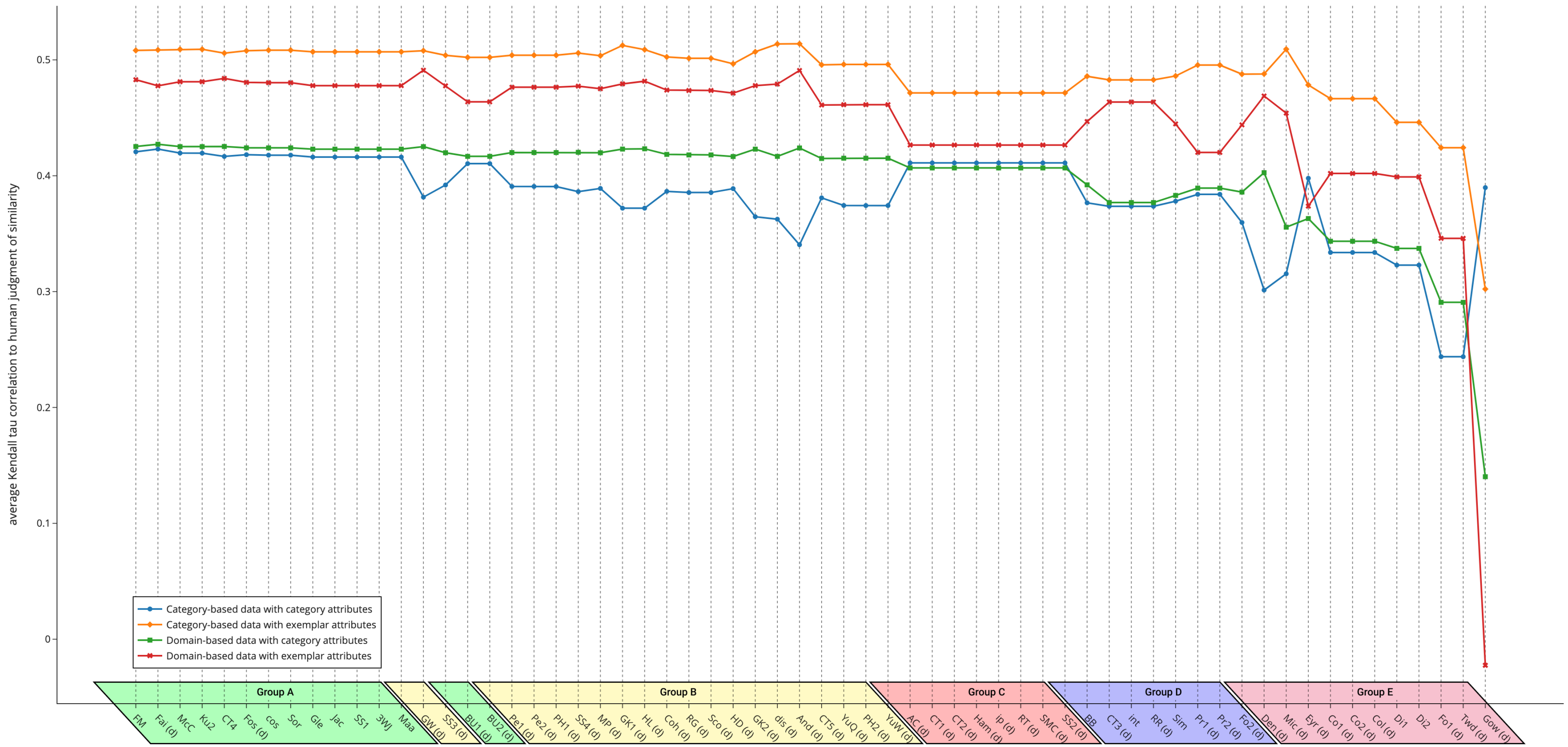

Figure 2: Average correlations of computed similarity to human judgment of similarity across categories of the artifact domain.

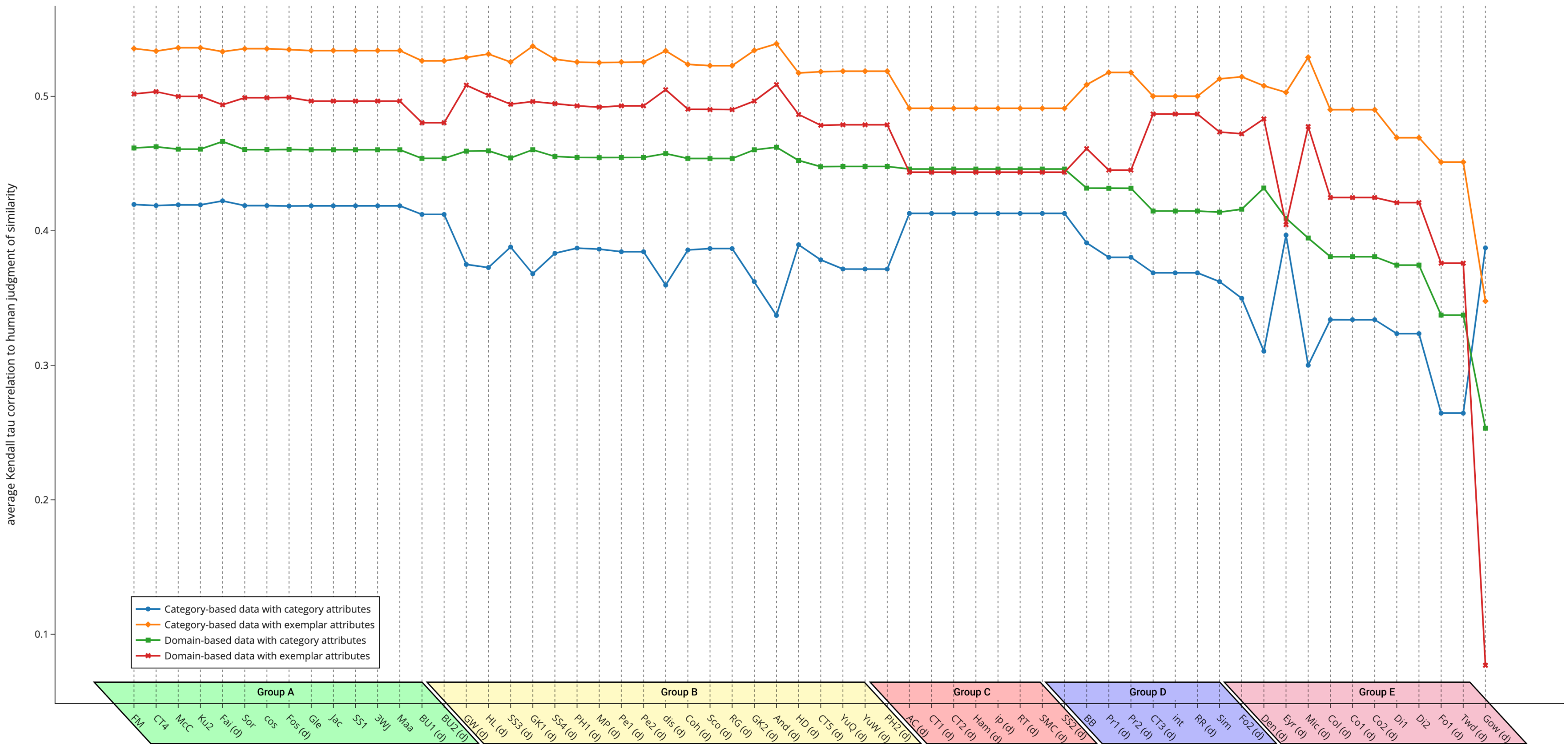

Figure 3: Average correlations of computed similarity to human judgment of similarity across across all data.

3.2.3. Discussion

Basic observations. Before analyzing the correlation graphs, note that according to a commonly accepted interpretation, the values of the Kendall τ_b correlation coefficient, which are displayed in our graphs, are interpreted as follows: $\tau_b \geq 0.3$, $0.2 \leq \tau_b < 0.3$, $0.1 \leq \tau_b < 0.2$, and $0.0 \leq \tau_b < 0.1$ indicate strong, moderate, weak, and very weak correlation, respectively.

Both the summarizing graphs as well as the detailed graphs in the online supplementary material demonstrate a notable fact, namely that except for a few cases, the examined similarity measures exhibit a strong correlation with the human judgment of similarity.[12] These correlations are naturally explained by the fact that each of the examined similarity measures has been designed to capture some aspects of human perception of similarity and indeed has proved useful in modeling similarity over the years. Yet, the rather high correlations with the human judgment of similarity across all of the examined categories appear remarkable in view of the fact that many of the similarity measures have been proposed to serve a particular purpose in a particular application domain. We also observed the corresponding p-values.[13] For the vast majority of the reported correlation values, the corresponding p-values are reasonably low; of the total 3224 observed p-values, only 22 were larger than 0.00001.

Role of category and exemplar attributes. In spite of the overall solid correlations, a detailed examination of the graphs reveals certain patterns and tendencies. While the degrees $\mathrm{HJ}(x_i, x_j)$ of the human judgment of exemplar-to-exemplar similarity only depend on the particular exemplars x_i and x_j of a chosen category, the degrees $sim(x_i, x_j)$ depend on the particular binary matrix that includes x_i and x_j, i.e., on the corresponding collection of attributes using which the values $sim(x_i, x_j)$ are computed. There are four kinds of these binary matrices and hence four kinds of attribute collections: The exemplar attributes involved in the small, category-based matrices, the exemplar attributes in the large, domain-based matrices, the category attributes in the category-based matrices, and the category attributes in the domain-based matrices. It is a rather intuitive psychological knowledge mentioned in the literature but not supported by quantitative experiments [37] that exemplar and category attributes provide different kinds of information about the concerned exemplars, which

[12] As explained at the end of section 3.2.1, we present our results for the Kendall rank-order coefficient, mainly due to its convenient statistical properties and intuitive appeal, Nevertheless, we also observed the Spearman rank-order correlation. This yielded somewhat larger correlation values but similar correlation patterns over the examined datasets.

[13] Note that the null hypothesis for the Kendall correlation test is that the coefficient equals 0, i.e., there is no correlation.

reflect the way these two types of attributes are generated by human respondents; see section 3.1. As a rule, the exemplar attributes, which are generated in response to the individual exemplars of a given category, provide a more distinctive information about the exemplars compared to the category attributes, which are generated in response to the category name. This intuitive knowledge is supported by our experiments.

Namely, the summarizing graphs in figs. 1–3 expose that for both the small, category-based matrices and the large, domain-based matrices, the exemplar attributes yield higher correlations with the human judgment of similarity than the category attributes. In more detail, the highest correlations with the human judgment are obtained for the category-based matrices with exemplar attributes, while the lowest correspond to the category-based matrices with category attributes, leaving in between the correlations of the domain-based matrices with exemplar followed by those of the domain-based matrices with category attributes. We hypothesize that the reason is to be sought in the above-mentioned quality of description of the exemplars by the respective attribute collections.

In particular, the category-based data with exemplar attributes seem to provide the best description in that the exemplar attributes are highly distinctive and, at the same time, all directly relate to the exemplars of the given category.[14] For a given category, such as "vehicle," the attributes of the domain-based data, on the other hand, contain many attributes that do not directly relate to the exemplars of this category, namely those generated as a response to the exemplars of other categories of the domain. As an example, the attribute *isSharp* is one of the exemplar attributes of the artifact domain because it has been generated in response to the exemplars of the category "kitchen utensil," e.g., in response to *knife*. This attribute is intuitively irrelevant for the exemplars of the category "vehicle." But since both "kitchen utensil" and "vehicle" are parts of the artifact domain, the attribute *isSharp* is actually used when computing the degrees of similarity of the exemplars of "vehicle" the same way as the attributes directly related to the exemplars of "vehicle", such

[14]Note in this connection that the Dutch data also provides information about reliability of the human judgment of similarity for each of the involved human categories, which offers the question of a relationship between this reliability and the examined Kendall correlations of the similarity measures with the human judgment of similarity. We hence explored rank-order correlation between the above-mentioned reliability for the involved the human categories on the one hand and the average of the Kendall correlation over all the similarity measures with human judgment of similarity for the human categories. For the presumably best description of exemplars by the attributes, i.e., the description of category-based data by exemplar attributes, the correlation indeed is the highest and turns out to be strong (Kendall tau slightly higher than 0.3).

as *usesFuel* or *hasWheels*. As a result, consideration of such irrelevant attributes in the domain data presents a bias and worsens the quality of the computed similarity degrees of the exemplars of "vehicle." The lowest correlations of similarity degrees were computed using the category-based matrices with category attributes are likely because the small number of them and the lack of their specificity result in insufficient information they carry about the exemplars.

Role of shared absences. Another topic, often discussed in the literature on the measures of similarity for binary data, is the role of the parameter d representing shared absences of attributes; see (1) in section 2. The point of debate in the literature consists in whether shared absences contribute to similarity. Some studies suggest that they do not contribute and, hence, that d should not be employed in the formulas for similarity measures. Some, however, claim that shared absences do contribute, albeit possibly to a smaller extent than the shared presences represented by the value a, and, hence, that d should be employed with a weight moderating the contribution of d. See, e.g., [7, 55] for details and references. In this respect, our experiments reveal that, as a rule, the measures not employing d yield stronger correlations with the human judgment of similarity; see the summarizing graphs in figs. 1–3, and hence appear to support the view that shared absences do not contribute to the perception of similarity.

Role of categories' kind. Another question, which is worth further examination from the psychology viewpoint, is whether the correlations with the human judgment of similarity are affected by the kind of categories. Overall, the correlations are higher for the natural categories than for the artifact ones, as is apparent from the graphs in figs. 1 and. 2. This is presumably due to a more peculiar nature of the artifact categories, which is known in the psychology of concepts. For instance, the artifact categories tend to be defined not only by their exemplars' attributes but to a great extent also by the goals of the exemplars of the category [12]. The attributes may hence carry less information about the exemplars of the artifact categories compared to the natural ones, hence the overall lower correlations. A closer look at the detailed graphs in the online supplementary material makes it also apparent that within each domain, the correlations of the degrees of similarity measures with the human judgment of similarity are consistently higher for some categories and consistently lower for others. In the animal domain, the categories with higher correlations include "fish," "reptile," and "bird," while those with lower correlations include "vegetable," "sport," and in data with the category attributes also "mammal." In the artifact domain, higher correlations appear consistently in particular for "weapon," and in data with the exemplar attributes also for "vehicle" and "musical instrument." While

AC, CT1, CT2, Ham, ip, RT, SMC, SS2
BU1, BU2
Co1, Co2
CT3, int, RR
Di1, Di2
Fo1, Twd
3WJ, Gle, Jac, Maa, SS1
Pe1, Pe2
Pr1, Pr2

Table 3: Classes of similarity measures that yield the same Kendall tau correlation.

these observations may indicate a more informative description by attributes of the exemplars for some categories, they also may be evidence of a varying ability of humans to assess the similarity of exemplars of certain categories.

Ordinal equivalence of similarity measures. Let us also observe that some of the measures of similarity, such as SMC and ip, AC and CT1, or CT3 and RR, yield the same values of Kendall correlation with the human judgment of similarity, even though the respective measures yield different degrees of similarity for the examined data. Table 3 displays the classes of the thus equivalent measures in that all the measures of any line of the table yield the same values of Kendall correlation for all of the examined data. Note that for some pairs of measures, such as SMC and ip, their equivalence is obvious from the corresponding formulas for computing the degrees of similarity. For some, however, this is not obvious, and the observed equivalence may in turn be due to the particular data used in our experiments, which renders a further analytic examination of the measures of similarity as an interesting topic for future exploration.[15] In this connection, let us also note that the other rank-order correlation we examined, the Spearman correlation, yields the same classes of equivalent measures.

Groups of similarity measures. The graph in fig. 3 summarizing correlations over all data suggests an intuitive partitioning of all the involved similarity measures into five groups, denoted A, B, C, D, and E, according to their performance in terms of Kendall correlation with the human judgment of similarity. This partitioning is indicated on

[15] [7] and [55] list different but related classes of equivalent measures which indicate that some of the perfect correlations are due to the particular data used to evaluate the measures.

the horizontal axes of all the summarizing graphs in figs. 1–3. Group A consists of the measures consistently exhibiting very high correlations. The correlations of all these measures are roundly the same and, hence, all these measures may be considered the best as regards agreement with a human judgment. The measures in group B fare quite well as well, but slightly worse than those of group A, particularly on the category-based matrices with category attributes. Group C includes measures that are indistinguishable by their correlations displayed on the examined data; see also table 3 and the discussion in the preceding paragraph. Group D is similar to group B in that the involved measures yield high correlations except for the category-based matrices with category attributes, and also in that their performance on the various kinds of matrices is slightly less consistent than for the measures of groups A and C. The measures in Group D may be thought of as the least consistent and worst performing. In particular, Gow even exhibits slightly negative correlations for some data.

Various groupings and structures of similarity measures have been suggested in the literature but [7] and [55] seem to be the only ones involving a large number of measures and having the identified structures based on extensive data. Interestingly, our groups naturally align with the structures identified in these two studies.

In particular, fig. 4 shows a two-dimensional plot of a multidimensional scaling obtained in [55],[16] in which our abbreviations for the measures are used as labels, and the coloring indicates our groups A–E (the colors are those used on the horizontal axis in figs. 1 and 2). This obvious alignment of our grouping with the result of the multidimensional scaling suggests fundamental relationships among the similarity measures.

Another example is apparent from fig. 5 which depicts a list of similarity measures from [55, figure 8] ordered from the best to the worst by global performance of the measures, indicated by the vertical bars in the graph, in the task of virtual screening evaluated on two large datasets involving large collections of chemical molecules.[17] Again, our abbreviations for the measures are used on the horizontal axis, and the coloring of measures indicates our groups A–E. Except for group E, the other groups align reasonably with the global performance of measures in that, by and large, the

[16]The two-dimensional scaling is obtained from a 44×44 matrix of the Pearson correlations of the 44 measures involved in the study; each correlation value was calculated using 100 000 data points for which the values of the corresponding pair of measures were computed.

[17]The two employed datasets, MDDR and WOMBAT, contain information about 102 540 and 138 127 molecules, respectively, which are represented by their binary fingerprints each consisting of 1024 bits. The similarity measures are employed in a similarity search which is a part of the screening procedure.

measures of every group appear, with some exceptions, as segments of the whole list. Interestingly, the measures of groups A and B, which exhibit the best agreement with human perception of similarity, appear as the best performing measures in the list.

Yet another grouping of similarity measures based on extensive data was obtained in [7]. The authors explored similarity measures in view of varying base rates, i.e., percentage of 1s, of the attributes appearing in the binary vectors. Using a K-median clustering, they identified seven subsets of measures such that all the measures in each subset belong to the same cluster for each base-rate setup. [18] The subsets include three large and four small ones plus several ungrouped singleton subsets. The nontrivial subsets are shown in table 4 (the numbers are just labels of these subsets, i.e., carry no other information such as performance of the measures). [19] The coloring of the measures, indicating our groups A–E, reveals that Subset 1 comprises mostly the measures in group B and some measures of groups E and D, and that with a few exceptions, Subset 2 and Subset 3 consist of most of the measures in group A and all measures in group C, respectively.

The alignment of our groups A–E with the groups in the above-mentioned studies is remarkable. While our groups are derived from an agreement of similarity measures with human perception of similarity, the latter groups were identified in rather different scenarios, using different data as well as different criteria. The relationships among the measures of similarity and possible groupings of the measures should hence further be explored to obtain a better insight into the behavior and systematic characterization of the measures.

[18] They used 15 levels of a base rate. For each level, the authors generated 100 000 binary vectors and obtained a corresponding 69×69 correlation matrix representing the relationship of the 69 measures involved. From each such matrix, a corresponding K-median clustering was obtained with $K = 2$.

[19] The measures of the original subsets not included in our study are omitted in table 4.

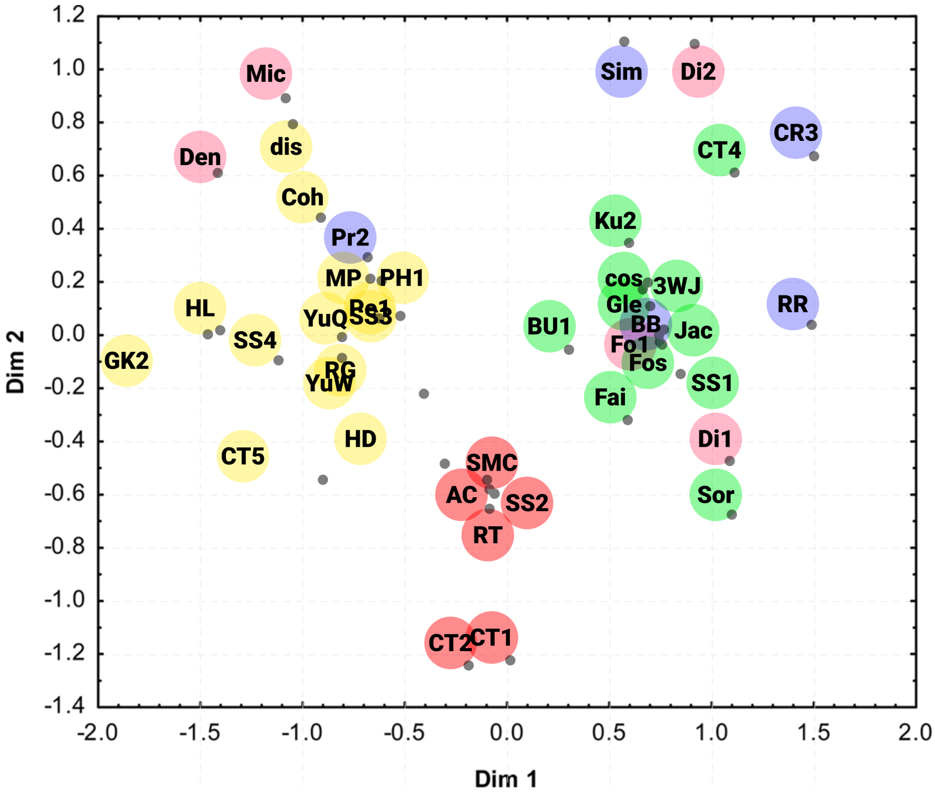

Figure 4: Multidimensional scaling of similarity measures of [55, figure 4] with colors indicating the groups A–E found in our study.

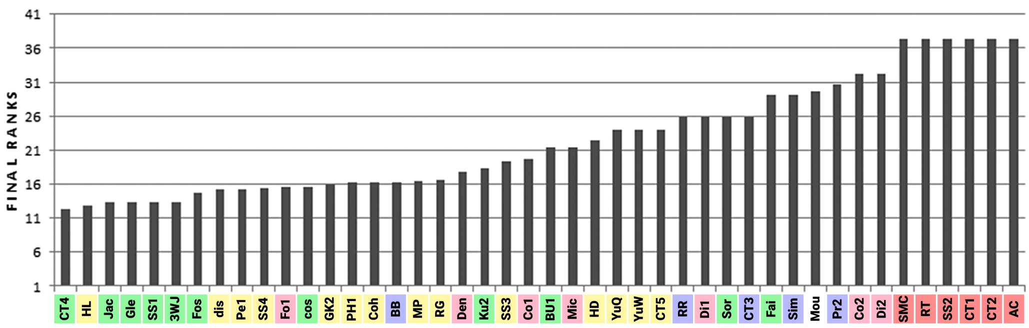

Figure 5: Global performance of similarity measures on large chemistry data from [55, figure 8] with colors indicating the groups A–E found in our study.

Subset 1	Co1	Co2	Coh	Fo1	Fo2	GW	MP	PH1	Pr1	Pr2	SS4	YuQ	YuW	dis
Subset 2	3WJ	BU1	BU2	CT3	Di2	Fos	Gle	Jac	SS1	Sor	cos			
Subset 3	AC	CT1	CT2	CT4	Fai	Ham	RT	SMC	SS2	SS3				
Subset 4	GK2	HL	RG	Sco										
Subset 5	Ku2	McC												
Subset 6	Pe1	Pe2												
Subset 7	Di1	Sim												

Table 4: Subsets of similarity measures identified in [7, table 3].

4. Conclusions

Measures of similarity represent a basic tool in processing binary data and have been explored for over a hundred years. While several comparative studies of similarity measures for binary data are available in the literature, these explore the properties of the measures and their mutual relationships but do not address the arguably basic question of agreement of the measures with human perception of similarity. To address this question, we identified a list of 69 well-known similarity measures that we employ in our study. We present this list along with references to the original sources we identified, and provide corrections of mistakes regarding properties of the existing measures that we found in the existing comparative papers and elsewhere in the literature. Our list enhances the other lists of similarity measures presented in the literature and may be used in further explorations.

To assess the agreement with human perception of similarity, we utilize now available high-quality Dutch data, which has been collected in a large-scale psychological study. In addition to several natural and artifact categories, the data includes exemplars in these categories, the human judgment of exemplar-to-exemplar similarity, various collections of attributes relevant to the included categories and exemplars, as well as the corresponding exemplar-attribute binary matrices. Using these binary matrices, we calculated the exemplar-by-exemplar similarity using the examined similarity measures and observed correlations with a human judgment of similarity. The most significant findings are the following:

- Except for a few cases, the examined similarity measures display overall rather strong correlations with a human judgment. While this appears to naturally follow from the fact that each of the examined measures has been designed to capture some aspects of human perception of similarity and has proved useful over the years, the rather high correlations with human judgment across all of the examined human categories are remarkable in view of the fact that many of the measures were proposed to serve a particular purpose in a particular domain.

- The revealed quality of correlation with human perception of similarity suggests five natural groups of similarity measures. The measures in these groups possess certain common general characteristics of similarity measures. Also, the groups are closely related to some classes of measures identified in studies that used synthetic and domain data rather than human judgment data to analyze measures of similarity.

- Correlations to a human judgment of exemplar similarity are higher when the computations of similarity degrees for the involved measures are based on the exemplar rather than the category attributes. This provides an experimental support for the intuitive knowledge, articulated without quantitative evidence in the literature on the psychology of concepts, that the exemplar attributes provide a more informative description of exemplars compared to the category attributes. Another observation of psychological relevance is that the correlations depend on the kind of categories to which the considered exemplars belong, and are generally higher for the natural categories and lower for the artifact ones.

Our study suggests several topics for further exploration. On a general note, the study accentuates the question of psychological plausibility of models used for information processing that are inspired by human cognition and a need for quality psychological data which are essential to explore this question. Even though the suitability of a concrete model may depend on the considered domain of application, the understanding of psychological plausibility helps to enhance our comprehension of the available models in a broader perspective, and to select the model in a given situation properly. As regards the measures of similarity for binary data, we consider the following three topics as particularly important. First, a further experimental study of similarity measures and their structure using both synthetic and real data to understand the behavior of the measures under various circumstances. The groups of measures revealed in our study and their significant alignment with the groups found in other studies, which we describe above, indicate deeper relationships among the measures. Second, a further theoretical exploration of the measures and the relationship among them. Even though various studies exist, an ordinal equivalence of measures indicated by the above-described classes of perfectly rank-order-correlated measures appears particularly interesting and has not been examined in the previous studies. Third, an investigation of psychological questions related to our experiments. This includes exploration of the descriptive ability of attributes and the question of category vs. exemplar attributes, particularly with respect to their ability to determine exemplar-to-exemplar similarity, as discussed above. Last but not least, assembly of quality datasets that would enable quantitative experimentation with measures of similarity and other methods tailored for binary data.

[1] Abydos: abydos.distance package. `https://abydos.readthedocs.io/en/latest/abydos.distance.html`.

[2] Anderberg, M. R.: Cluster Analysis for Applications. Academic Press, New York and London (1973).

[3] Austin, B., Colwell, R. R.: Evaluation of some coefficients for use in numerical taxonomy of microorganisms. International Journal of Systematic Microbiology 27, no. 3, 204–210 (1977).

[4] Baroni-Urbani, C., Buser, M. W.: Similarity of binary data. Systematic Zoology 25, no. 3, 251–259 (1976).

[5] Belohlavek, R., Mikula, T.: Dutch Concepts. GitHub repository (2023). `https://github.com/mikulatomas/dutch-concepts`.

[6] Braun-Blanquet, J.: Plant Sociology: The Study of Plant Communities. McGraw-Hill, New York (1932).

[7] Brusco, M., Cradit, J. D., Steinley, D.: A comparison of 71 binary similarity coefficients: the effect of base rates. PLOS ONE 16, no. 4, 1–19 (2021).

[8] Choi, S. S., Cha, S. H., Tappert, C. C.: A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics 8, no. 1, 43–48 (2010).

[9] Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, no. 1, 37–46 (1960).

[10] Cole, L. C.: The measurement of interspecific association. Ecology 30, no. 4, 411–424 (1949).

[11] Consonni, V., Todeschini, R.: New similarity coefficients for binary data. MATCH Communications in Mathematical and Computer Chemistry 68, no. 2, 581–592 (2012).

[12] De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., Storms, G.: Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. Behavior Research Methods 40, no. 4, 1030–1048 (2008).

[13] Dice, L. R.: Measures of the amount of ecologic association between species. Ecology 26, no. 3, 297–302 (1945).

[14] Driver, H. E., Kroeber, A. L.: Quantitative expression of cultural relationships. The University of California Publications in American Archaeology and Ethnology 31, no. 4, (1932).

[15] Eyraud, H.: Les principes de la mesure des corrélations. Annales de l'Université de Lyon, Série III, Section A, no. 1, 30–47 (1936).

[16] Faith, D. P.: Asymmetric binary similarity measures. Oecologia 57, 287–290 (1983).

[17] Fager, E. W.: Determination and analysis of recurrent groups. Ecology 38, no. 4, 586–595 (1957).

[18] Fager, E. W., McGowan, J. A.: Zooplankton species groups in the North Pacific: Co-occurrences of species can be used to derive groups whose members react similarly to water-mass types. Science 140, no. 3566, 453–460 (1963).

[19] Forbes, S. A.: On the local distribution of certain Illinois fishes: An essay in statistical ecology. Bulletin of the Illinois State Laboratory of Natural History 7, Article 8 (1907).

[20] Forbes, S. A.: Method of determining and measuring the associative relations of species. Science 61, no. 1585, 524 (1925)

[21] Fossum, E. G., Kaskey, G.: Optimization and standardization of information retrieval language and systems. Technical Report, Directorate of Information Sciences, Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, Washington, D.C., vi+87 pp. (1966).

[22] Gilbert, N., Wells T. C. E.: Analysis of quadrat data. The Journal of Ecology 54, no. 3, 675–685 (1966).

[23] Goodman, L. A., Kruskal, W. H.: Measures of association for cross classifications. J. American Statistical Association 49, no. 268, 732–764 (1954)

[24] Gower, J. C., Legendre, P.: Metric and Euclidean properties of dissimilarity coefficients. J. Classification 3, 5–48 (1986).

[25] Hamann, U.: Merkmalsbestand und Verwandtschaftsbeziehungen der farinosae: ein beitrag zum system der monokotyledonen. Willdenowia Bd. 2, H. 2, 639–768 (1961).

[26] Harris, F. C., Lahey, B. B.: A method for combining occurrence and nonoccurrence interobserver agreement scores. Journal of Applied Behavior Analysis 11, no. 4, 523–527 (1978).

[27] Hawkins R. P.; Dotson V. A.: Reliability scores that delude: An Alice in wonderful trip through the misleading characteristics of interobserver agreement scores in interval coding. Behavior Analysis: Areas of Research and Application, Prentice Hall, Englewood Cliffs, NJ (1968).

[28] Holliday, J. D., Hu, C. Y., Willett, P.: Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Combinatorial Chemistry & High Throughput Screening 5, no. 2, 155–166 (2002).

[29] Hubálek, Z.: Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biological Reviews 57, no. 4, 669–689 (1982).

[30] Jaccard, P.: Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bulletin de la Société des Sciences Naturelles 37, 241–272 (1901).

[31] Johnson, S. C.: Hierarchical clustering schemes. Psychometrika 32, no. 3, 241–254 (1967).

[32] Jones, J. A., Harrold, M. J., Stasko, J.: Visualization of test information to assist fault localization. In Proceedings of the 24th international conference on Software engineering, 467–477, Orlando, Florida (2002).

[33] Kulczynski, S.: Zespoly rostlin w Pieninach. Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles, Série B (Sciences Naturelles), Supplement II, 57–203 (1927).

[34] Maxwell, A. E., Pilliner, A. E. G.: Deriving coefficients of reliability and agreement for ratings. British Journal of Mathematical and Statistical Psychology 21, no. 1, 105–116 (1968).

[35] McConnaughey, B. H.: The determination and analysis of plankton communities. Marine Research in Indonesia Spec., 40 (1964).

[36] Michael, E. L.: Marine ecology and the coefficient of association: A plea in behalf of quantitative biology. Journal of Animal Ecology 8, no. 1, 54–59 (1920).

[37] Murphy, G. L.: The Big Book of Concepts. MIT Press, Cambridge, Mass. (2002).

[38] Mountford, M. D.: An index of similarity and its application to classificatory problems. In Murphy P. W. (Ed.), Progress in soil zoology, London: Butterworths, 43–50 (1962).

[39] Pearson K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5, 50:302, 157–175 (1900).

[40] Pearson, K., Heron, D.: On theories of association. Biometrika 9, no. 1/2, 159–315 (1913).

[41] Peirce, C. S.: The numerical measure of the success of predictions. Science 93, 453–454 (1884).

[42] Rogers, D. J., Tanimoto, T. T.: A computer program for classifying plants. Science 132, no. 3434, 1115–1118 (1960).

[43] Rogot, E., Goldberg, I. D.: A proposed index for measuring agreement in test-retest studies. Journal of Chronic Diseases 19, no. 9, 991–1006 (1966).

[44] Rosch, E.: Principles of categorization. In Rosch, E., Lloyd, B. B. (Eds.): Cognition and Categorization, Erlbaum, Hillsdale, NJ, 27–48 (1978).

[45] Russell, P. F., Rao, T. R.: On habitat and association of species of anopheline larvae in south-eastern Madras. Journal of the Malaria Institute of India 3, no. 1, 153–178 (1940).

[46] Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., Storms, G.: Dutch norm data for 13 semantic categories and 338 exemplars. Behavior Research Methods, Instruments, & Computers 36, no. 3, 506–515 (2004).

[47] Scott, W. A.: Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly 19, no. 3, 321–325 (1955).

[48] Simpson, G. G.: Mammals and the nature of continents. American Journal of Science 241, no. 1, 1–31 (1943).

[49] Sokal, R. R., Michener, C. D.: A statistical method for evaluating systematic relationships. The University of Kansas Science Bulletin 38, part 2, no. 22, 1409–1438 (1958).

[50] Sokal, R. R., Sneath P. H. A.: Principles of Numerical Taxonomy. W. H. Freeman and Company, San Francisco (1963).

[51] Sorensen, T. J.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biologiske skrifter 5, 1–34 (1948).

[52] Sorgenfrei, T.: Molluscan assemblages from the marine middle Miocene of South Jutland and their environments. Danmarks Geologiske Undersøgelse II. Række, no. 79, 403 (1958).

[53] Stiles, H. E.: The association factor in information retrieval. Journal of the ACM 8, no. 2, 271–279 (1961).

[54] Tarwid, K.: Szacowanie zbieznosci nisz ekologicznych gatunkow droga oceny prawdopodobienstwa spotykania sie ich w polowach. Ekologia Polska, Seria B, 115–130 (1960).

[55] Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., Willett, P.: Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. Journal of Chemical Information and Modeling 52, no. 11, 2884–2901 (2012).

[56] van der Maarel, E.: On the use of ordination models in phytosociology. Vegetatio 19, 21–46 (1969).

[57] Virtanen, P., Gommers, R., Oliphant, T. E. et al.: SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020).

[58] Warrens, M. J.: Similarity Coefficients for Binary Data. PhD thesis, Leiden University, The Netherlands (2008). `https://scholarlypublications.universiteitleiden.nl/access/item%3A2961378/view`.

[59] Similarity Coefficients for Binary Data

[60] Yule, G. U.: On the methods of measuring association between two attributes. Journal of the Royal Statistical Society 75, no. 6, 579–652 (1912).

Appendix: Table of similarity measures

symbol	name	formula	source
AC	Austin-Colwell	$\frac{2}{\pi}\arcsin\sqrt{\frac{a+d}{n}}$	[3]
And	Anderberg	$\frac{\tau_1-\tau_2}{2n}$ with $\begin{array}{l}\tau_1=\max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)\\ \tau_2=\max(a+c,b+d)+\max(a+b,c+d)\end{array}$	[2]
BB	Braun-Blanquet	$\frac{a}{\max(a+b,a+c)}$	[6]
BU1	Baroni-Urbani-Buser 1	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	[4]
BU2	Baroni-Urbani-Buser 2	$\frac{\sqrt{ad}+a-b-c}{\sqrt{ad}+a+b+c}$	[4]
Coh	Cohen	$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	[9]
Col	Cole	$\begin{cases}\frac{ad-bc}{(a+b)(b+d)} & \text{if } ad \geq bc\\ \frac{ad-bc}{(a+b)(a+c)} & \text{if } ad < bc \text{ and } d \geq a\\ \frac{ad-bc}{(b+d)(c+d)} & \text{otherwise}\end{cases}$	[10]
Co1	Cole (Cole 1)	$\frac{ad-bc}{(a+c)(c+d)}$	[10]
Co2	Cole (Cole 2)	$\frac{ad-bc}{(a+b)(b+d)}$	[10]
cos	cosine (Driver-Kroeber, Ochiai)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[14]
CT1	Consonni-Todeschini 1	$\frac{\ln(1+a+d)}{\ln(1+n)}$	[11]
CT2	Consonni-Todeschini 2	$\frac{\ln(1+n)-\ln(1+b+c)}{\ln(1+n)}$	[11]
CT3	Consonni-Todeschini 3	$\frac{\ln(1+a)}{\ln(1+n)}$	[11]
CT4	Consonni-Todeschini 4	$\frac{\ln(1+a)}{\ln(1+a+b+c)}$	[11]
CT5	Consonni-Todeschini 5	$\frac{\ln(1+ad)-\ln(1+bc)}{\ln(1+n^2/4)}$	[11]
Den	Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	*[28]

dis	dispersion	$\frac{ad-bc}{n^2}$	*[8]
Di1	Dice 1	$\frac{a}{a+b}$	[13]
Di2	Dice 2	$\frac{a}{a+c}$	[13]
Eyr	Eyraud	$\frac{n^2(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)}$	[15]
Fai	Faith	$\frac{a+0.5d}{n}$	[16]
FM	Fager-McGowan	$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{\max(a+b,a+c)}}$	[18]
Fos	Fossum	$\frac{n(a-\frac{1}{2})^2}{(a+b)(a+c)}$	[21]
Fo1	Forbes 1	$\frac{na}{(a+b)(a+c)}$	[19]
Fo2	Forbes 2	$\frac{na-(a+b)(a+c)}{n\min(a+b,a+c)-(a+b)(a+c)}$	[20]
Gle	Gleason (Dice, Sørensen, Czekanowski)	$\frac{2a}{2a+b+c}$	*[13, 51]
GK1	Goodman-Kruskal 1	$\frac{\tau_1-\tau_2}{2n-\tau_2}$ with $\tau_1 = \max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)$, $\tau_2 = \max(a+c,b+d)+\max(a+b,c+d)$	[23]
GK2	Goodman-Kruskal 2	$\frac{2\min(a,d)-b-c}{2\min(a,d)+b+c}$	[23]
Gow	Gower	$\frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	*[8]
GW	Gilbert-Wells	$\ln\frac{n^3}{2\pi(a+b)(c+d)(a+c)(b+d)} + 2\ln\frac{n!a!b!c!d!}{(a+b)!(c+d)!(a+c)!(b+d)!}$	[22]
Ham	Hamman	$\frac{a+d-b-c}{n}$	[25]
HD	Hawkins-Dotson	$\frac{1}{2}\left(\frac{a}{a+b+c}+\frac{d}{d+b+c}\right)$	[27]
HL	Harris-Lahey	$\frac{a(2d+b+c)}{2(a+b+c)}+\frac{d(2a+b+c)}{2(b+c+d)}$	[26]
int	intersection	a	*[8]
ip	inner product	$a+d$	*[8]

Jac	Jaccard (Jaccard-Tanimoto)	$\frac{a}{a+b+c}$	[30]
Ku1	Kulczynski 1	$\frac{a}{b+c}$	[33]
Ku2	Kulczynski 2 (Driver-Kroeber)	$\frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c})$	[33]
Maa	van der Maarel	$\frac{2a-b-c}{2a+b+c}$	[56]
McC	McConnaughey	$\frac{a^2-bc}{(a+b)(a+c)}$	[35]
Mic	Michael	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	[36]
Mou	Mountford	$\frac{2a}{ab+ac+2bc}$	[38]
MP	Maxwell-Pilliner	$\frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	[34]
Pe1	Pearson 1 (χ^2 statistical significance)	$\frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$	[39]
Pe2	Pearson 2	$\sqrt{\frac{\chi^2}{n+\chi^2}}$ with χ^2 equal to Pe1	[39]
Pe3	Pearson 3	$\sqrt{\frac{\rho}{n+\rho}}$ with ρ equal to PH1	*[8]
PH1	Pearson-Heron 1 (Phi)	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	[40]
PH2	Pearson-Heron 2	$\cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)$	[40]
Pr1	Peirce 1	$\frac{ad-bc}{(a+b)(c+d)}$	[41]
Pr2	Peirce 2	$\frac{ad-bc}{(a+c)(b+d)}$	[41]
Pr3	Peirce 3	$\frac{ad+bc}{ab+2bc+cd}$	*[29]
RG	Rogot-Goldberg	$\frac{a}{2a+b+c} + \frac{d}{2d+b+c}$	[43]
RR	Russel-Rao	$\frac{a}{n}$	[45]
RT	Rogers-Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	[42]

Sco	Scott	$\frac{4ad-(b+c)^2}{(2a+b+c)(2d+b+c)}$	[47]
Sim	Simpson	$\frac{a}{\min(a+b,a+c)}$	[48]
SMC	simple matching coefficient (Sokal-Michener)	$\frac{a+d}{n}$	[49]
Sor	Sorgenfrei	$\frac{a^2}{(a+b)(a+c)}$	[52]
SS1	Sokal-Sneath 1	$\frac{a}{a+2b+2c}$	[50]
SS2	Sokal-Sneath 2	$\frac{2a+2d}{2a+b+c+2d}$	[50]
SS3	Sokal-Sneath 3	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d})$	[50]
SS4	Sokal-Sneath 4, Ochiai 2	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	[50]
SS5	Sokal-Sneath 5	$\frac{a+d}{b+c}$	[50]
Sti	Stiles	$\log_{10}\frac{n(\lvert an-bc\rvert-\frac{1}{2}n)^2}{bc(n-b)(n-c)}$	[53]
Tar	Tarantula	$\frac{a(c+d)}{c(a+b)}=\frac{\frac{a}{a+b}}{\frac{c}{c+d}}$	[32]
Twd	Tarwid	$\frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$	[54]
YuQ	Yule (Yule Q)	$\frac{ad-bc}{ad+bc}$	[60]
YuW	Yule (Yule W)	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	[60]
3WJ	3W-Jaccard	$\frac{3a}{3a+b+c}$	[30]

Table 5: Measures of similarity. See section 2.2 for description of the columns.

Online supplementary material

The following eight figures, fig. 6–fig. 13, are to be published as the online supplementary material per journal style.

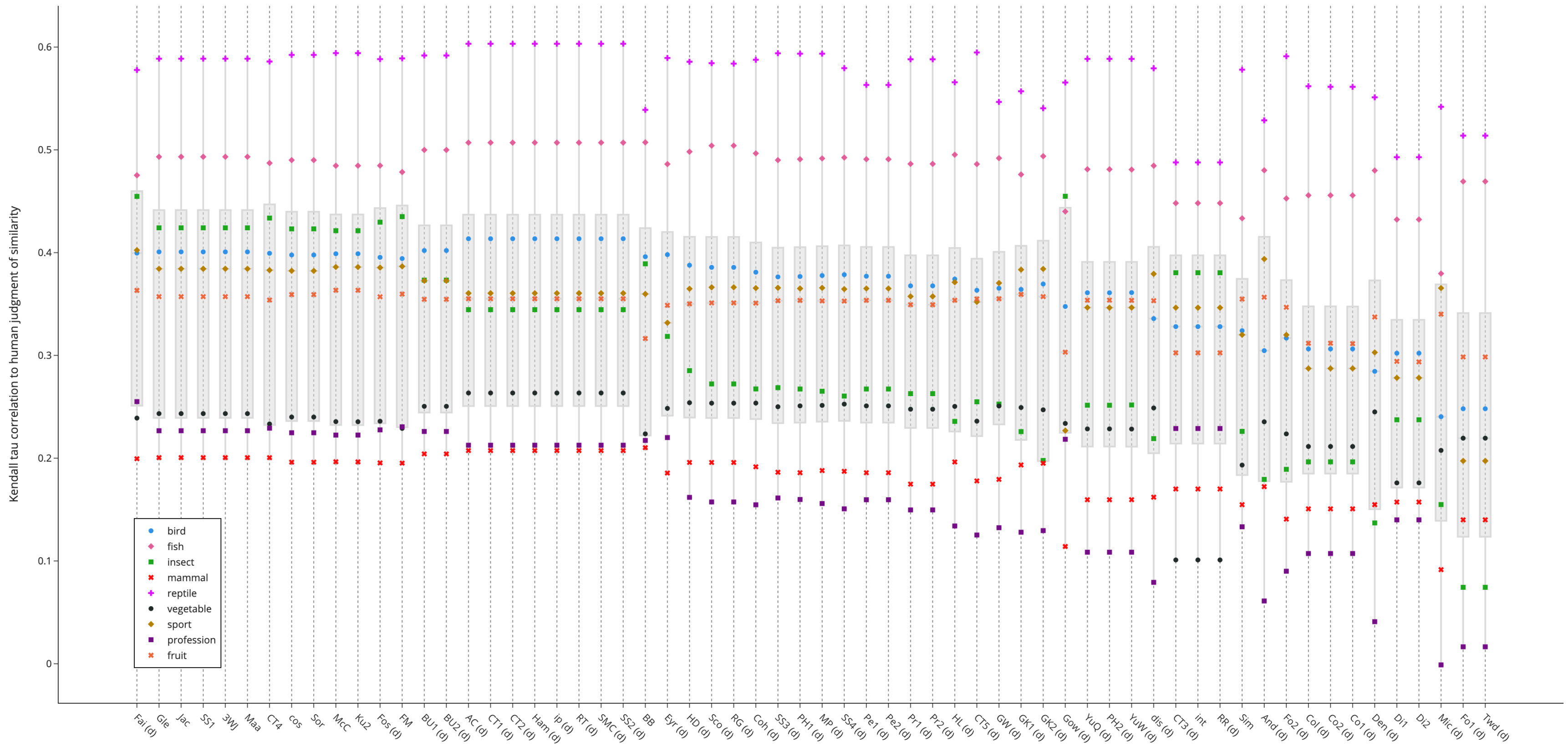

Figure 6: Correlations of computed similarity to human judgment of similarity across natural categories with category attributes (horizontal axis ordered by mean value).

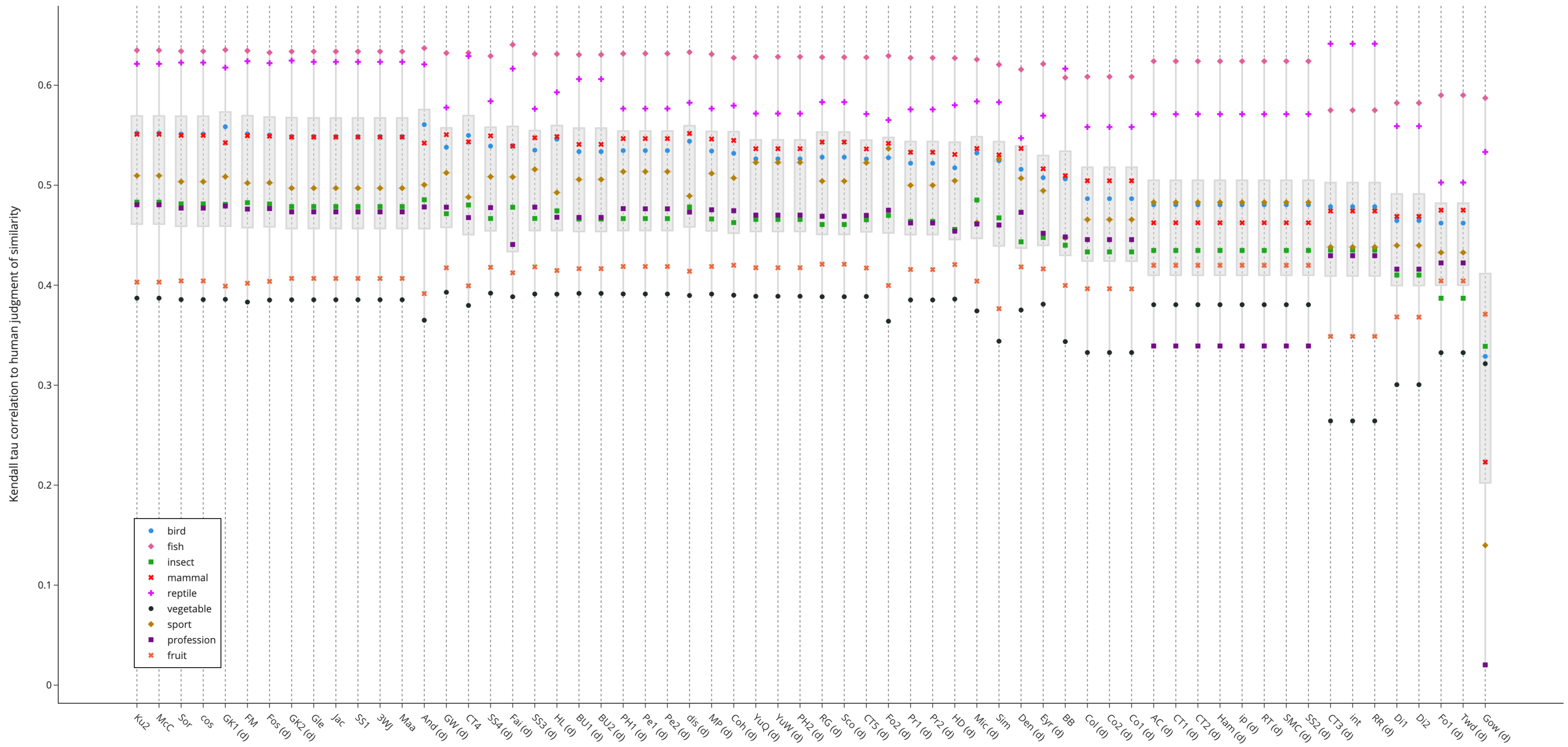

Figure 7: Correlations of computed similarity to human judgment of similarity across natural categories with exemplar attributes (horizontal axis ordered by mean value).

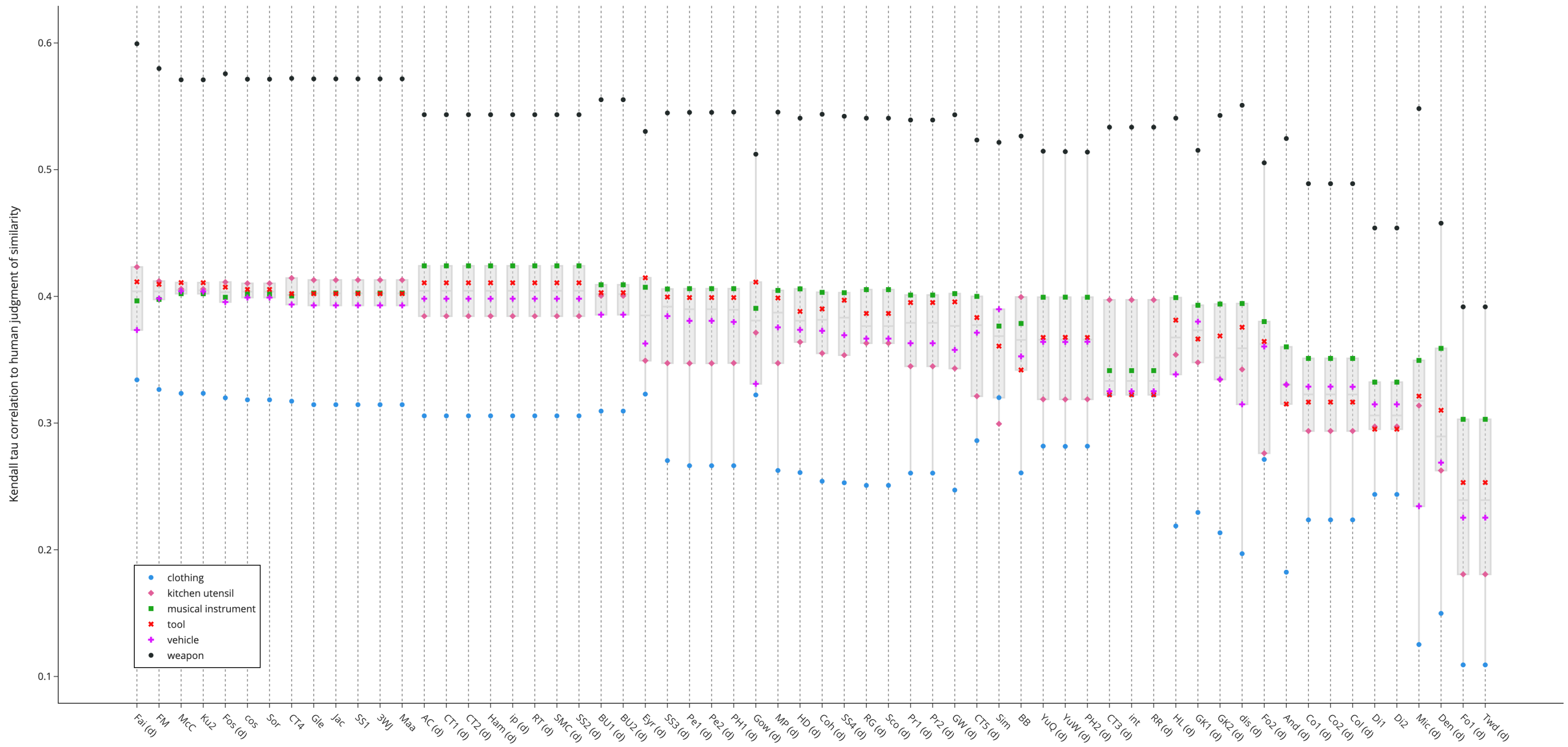

Figure 8: Correlations of computed similarity to human judgment of similarity across artifact categories with category attributes (horizontal axis ordered by mean value).

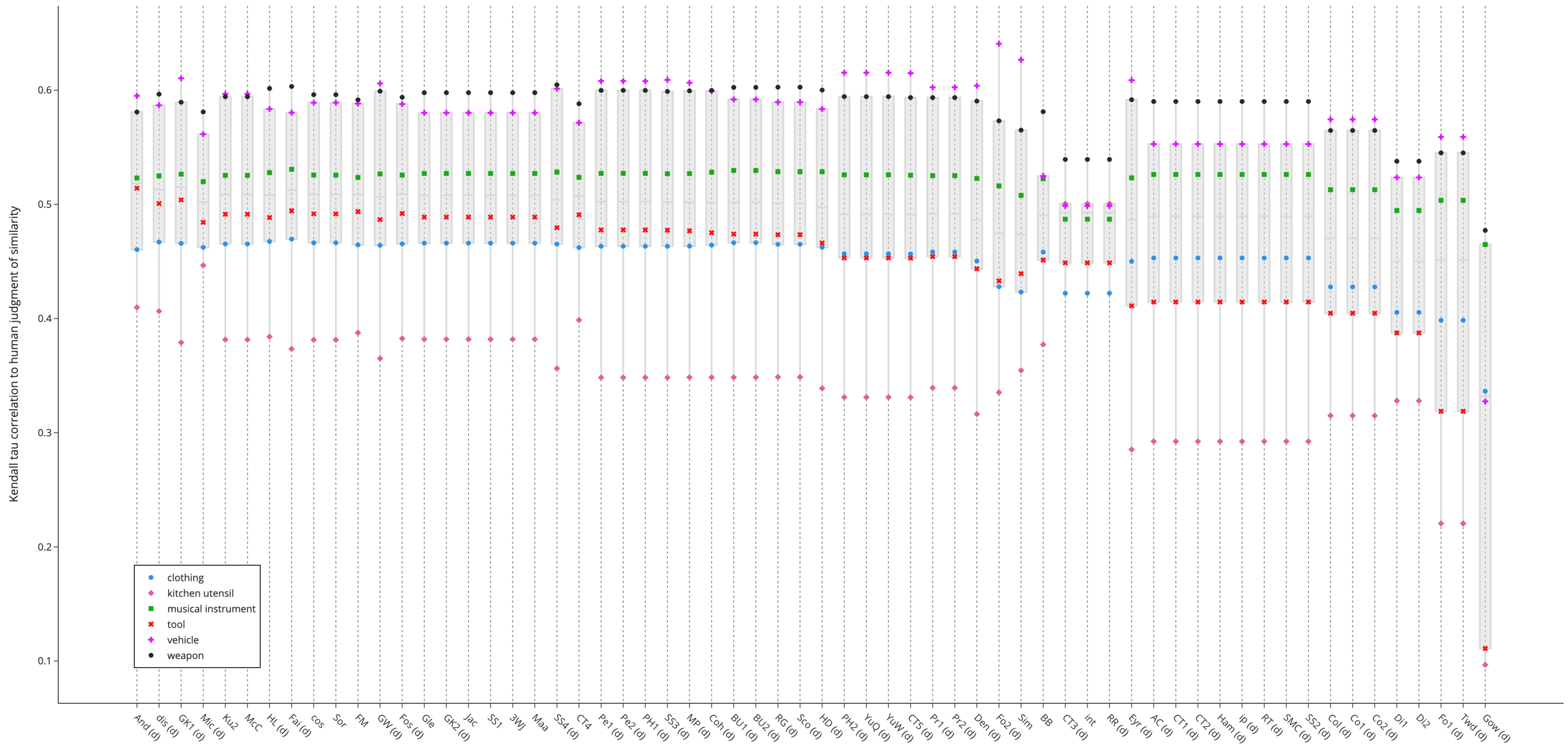

Figure 9: Correlations of computed similarity to human judgment of similarity across artifact categories with exemplar attributes (horizontal axis ordered by mean value).

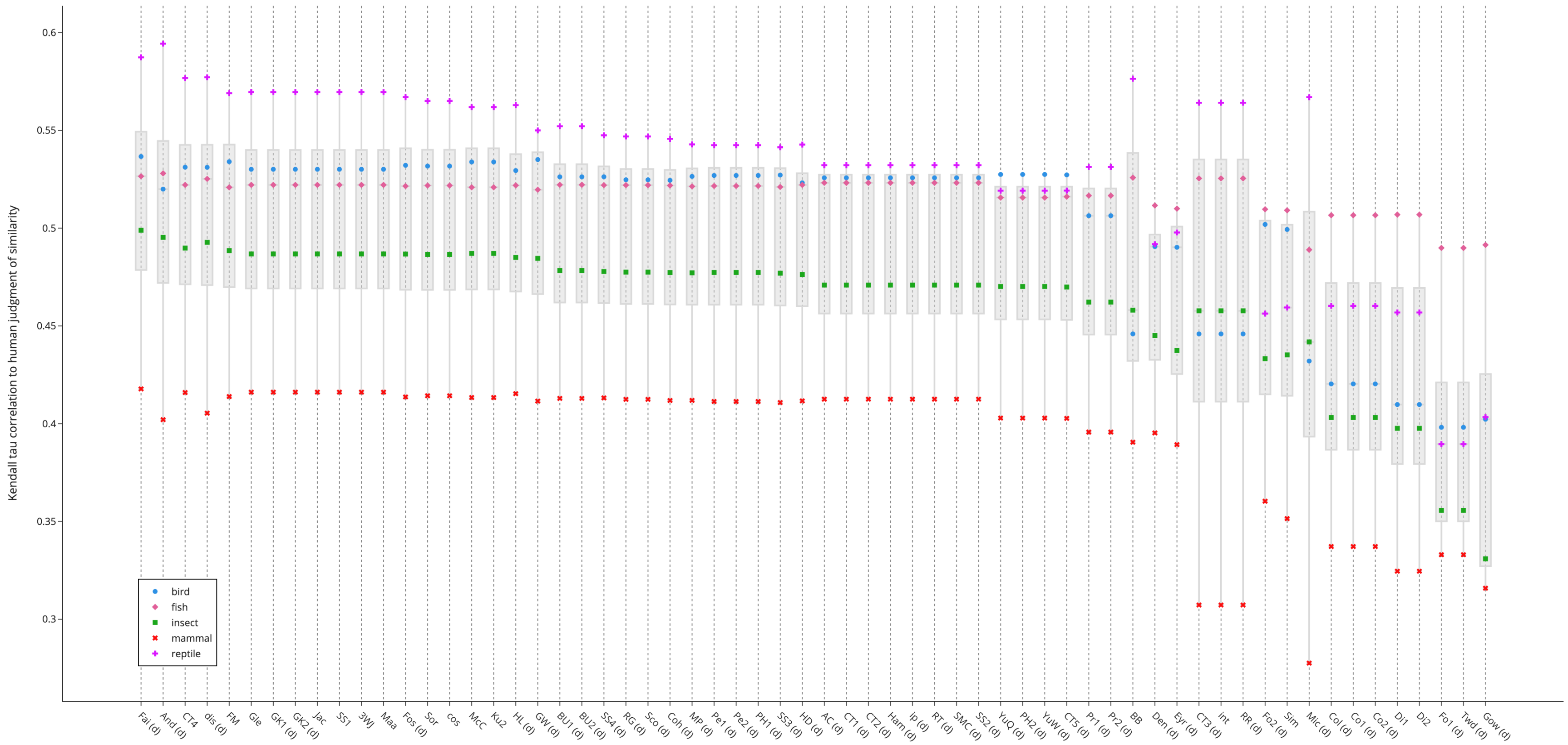

Figure 10: Correlations of computed similarity to human judgment of similarity across animal domain with category attributes (horizontal axis ordered by mean value).

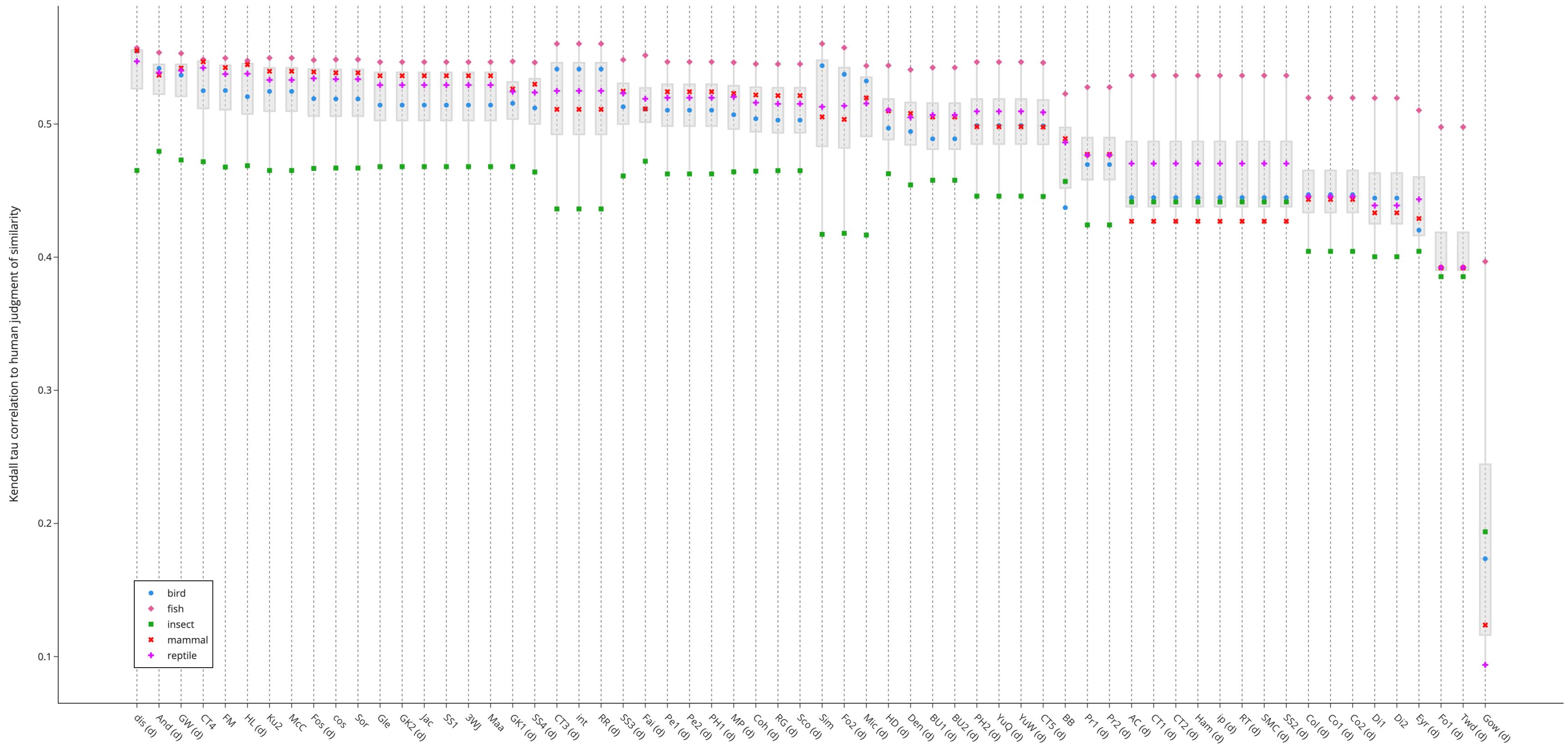

Figure 11: Correlations of computed similarity to human judgment of similarity across animal domain with exemplar attributes (horizontal axis ordered by mean value).

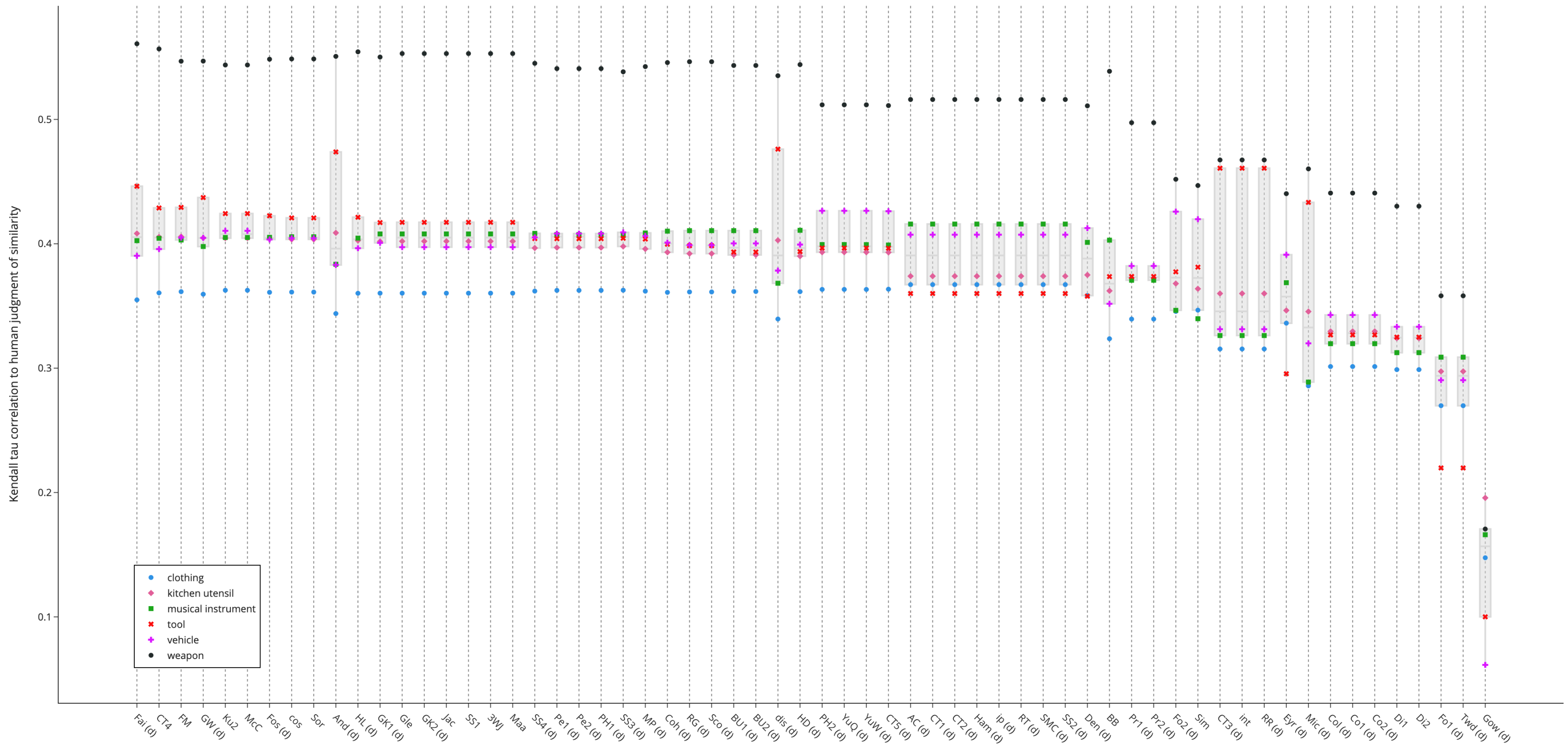

Figure 12: Correlations of computed similarity to human judgment of similarity across artifact domain with category attributes (horizontal axis ordered by mean value).

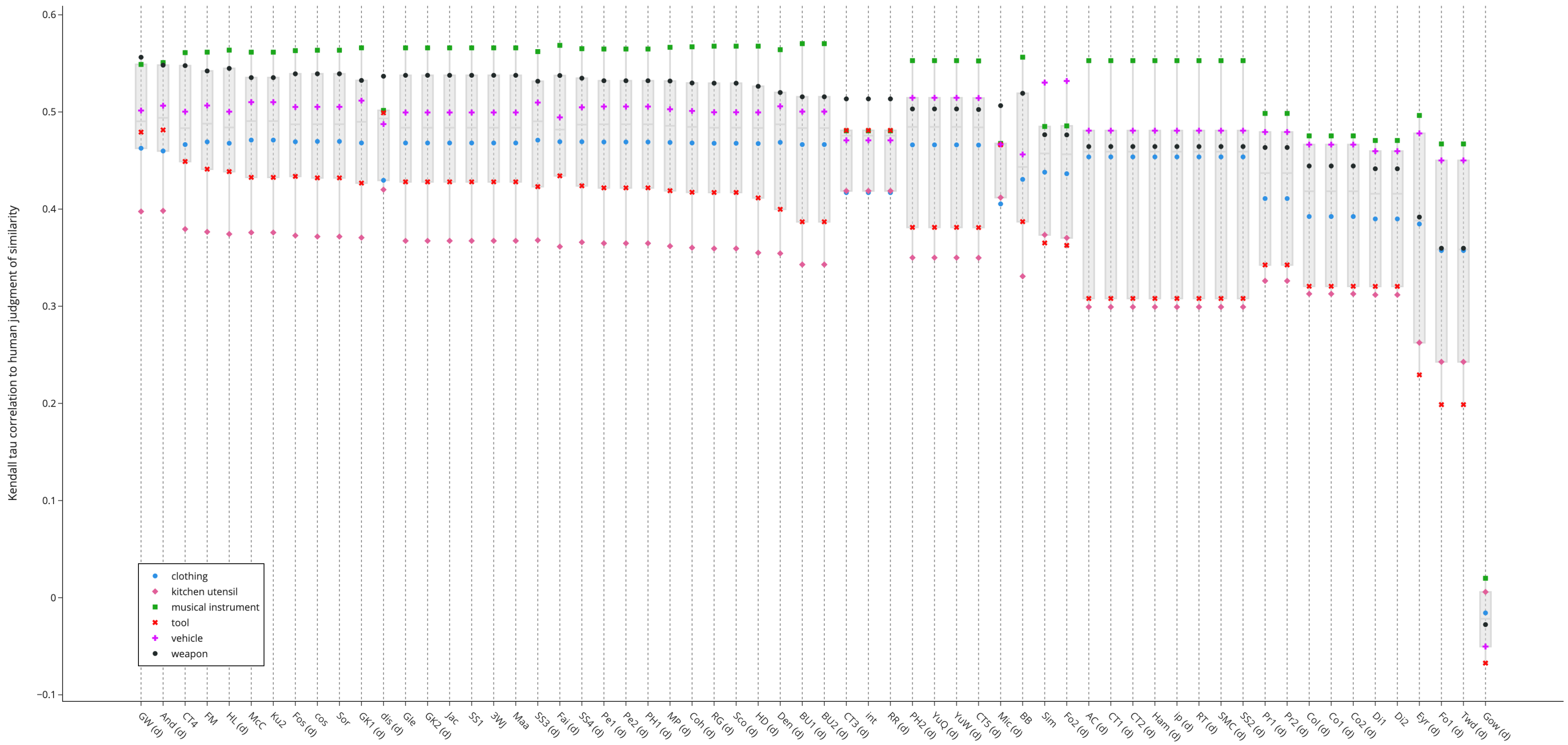

Figure 13: Correlations of computed similarity to human judgment of similarity across artifact domain with exemplar attributes (horizontal axis ordered by mean value).

Appendix E

Are human categories formal concepts? A case study using Dutch data

In this paper (Belohlavek and Mikula, 2024a), we examine to which extent the categories from the Dutch data form formal concepts. The results of these experiments are briefly described in Section 4.6. The paper resulted from joint research with my supervisor Radim Bělohlávek and was published in the International Journal of General Systems.

INTERNATIONAL JOURNAL OF GENERAL SYSTEMS
https://doi.org/10.1080/03081079.2024.2311917

Are human categories formal concepts? A case study using Dutch data

Radim Belohlavek and Tomas Mikula

Department of Computer Science, Palacký University Olomouc, Olomouc, Czech Republic

ABSTRACT

While the notion of a formal concept, as used in formal concept analysis, is inspired by the traditional view of human concepts, the psychological relevance of formal concepts has not been examined in the past. In this paper, we provide an experimental exploration of the psychological plausibility of formal concepts as human categories. For this purpose, we use the currently most extensive available psychological data regarding human categories. The data involve several human categories, over 400 exemplars of these categories, several hundreds of binary attributes that describe these exemplars and several binary matrices representing which exemplars have which attributes. Our primary question is: Are human categories formal concepts? That is, do the involved human categories represent formal concepts in the respective exemplar-attribute binary matrices? In most of the examined instances, the answer to this question turns out affirmative. This supports the hypothesis that formal concepts provide a psychologically plausible model of human categories. In addition, we discuss several related questions, provide observations on the psychological data used and present topics for future exploration.

ARTICLE HISTORY
Received 26 August 2023
Accepted 24 January 2024

KEYWORDS
Category; concept; formal concept; psychology; data analysis

1. Human concepts versus formal models of concepts and our aim

Concepts, or categories, are central to human reasoning (Machery 2007; Murphy 2002; Smith and Medin 1981). Various attempts have been made to provide formal models of concepts and concept formation and include numerous approaches in logic, machine learning, data mining and artificial intelligence. These have been studied mainly from the mathematical and computational viewpoints, and the viewpoint of possible applications in various domains.

A viewpoint central to our paper is that of a psychological relevance of formal models of concepts. Clearly, this viewpoint is of great importance not only from the psychological perspective but also from the viewpoint of the overall value of the particular model of concepts. Yet, studies of the psychological relevance of formal models of concepts are

CONTACT Tomas Mikula tomas.mikula01@upol.cz

virtually non-existent. The rare comments on the psychological relevance are restricted to intuitive considerations, unsubstantiated by experiments with solid psychological data. While the general reason for this unsatisfactory situation is, for the most part, the ignorance of the question of psychological relevance in the respective fields mentioned above, a particular obstacle in a possible pursuit of this question is a lack of proper psychological data that would make the pursuit realizable.

The main aim of our paper is to explore experimentally the psychological plausibility of the notion of a formal concept – a well-known simple mathematical model of human concepts employed in formal concept analysis (Carpineto and Romano 2004; Ganter and Wille 1999). The notion of a formal concept is based on a long-standing, traditional understanding of human concepts, worked out in particular in the Port-Royal logic (Arnauld and Nicole 1962), as an entity consisting of its extent, i.e. a collection of objects to which the concept applies, and its intent, i.e. a group of attributes characteristic of the concept. From the viewpoint of the psychology of concepts, formal concept analysis may be regarded as a simple formalization of an old tradition, known as the classical view of concepts (Murphy 2002; Smith and Medin 1981), according to which a human concept (category) is determined by a collection of its defining attributes (characteristics): An object is a member of the category if and only if it has each of the defining attributes.

Formal concepts proved useful in various domains; see, e.g. the books by Carpineto and Romano (2004), Ganter et al. (2016) and Ganter, Stumme, and Wille (2005). The corresponding mathematical and computational foundations are well developed and are still subject of current research (Ganter et al. 2016; Ganter and Wille 1999).[1] Still the basic question of whether the notion of a formal concept is psychologically plausible has not been studied in the past. The above-described situation applies also in a broader sense in that with some exceptions, the psychological relevance of the notions and results in formal concept analysis is ignored.[2]

We approach the question of psychological plausibility of formal concepts experimentally using the now available Dutch data (De Deyne et al. 2008; Ruts et al. 2004); see also Belohlavek and Mikula (2023). The Dutch data represents high-quality psychological data concerning human categories, which was gathered from several hundred respondents (De Deyne et al. 2008; Ruts et al. 2004). The data is particularly suited for our purpose as it involves a variety of human categories with a broad coverage, hundreds of exemplars of these categories and hundreds of attributes, i.e. features, pertaining to these categories and exemplars, along with a number of binary matrices representing which exemplars have which features. Our most significant finding is that in most instances, the human categories of the Dutch data indeed do form formal concepts in the respective binary matrices. In addition, we provide various related observations regarding formal concepts as well as the Dutch data itself.

Our paper is organized as follows. In Section 2, we provide preliminaries on formal concept analysis and, in particular, present the notion of a formal concept, provide relevant information from the psychology of concepts and describe the Dutch data. In Section 3, we present our experimental evaluation, observations and a discussion of the experimental results. Concluding remarks and topics for future exploration are the content of Section 4.

Table 1. Binary matrix representing five objects (rows), four attributes (columns), and a relation I between objects and attributes.

I	y_1	y_2	y_3	y_4
x_1	0	1	1	0
x_2	0	1	1	0
x_3	0	1	1	1
x_4	1	1	1	1
x_5	1	0	0	0

2. Formal concepts and the Dutch data

2.1. Formal concepts

The notion of a formal concept is the central notion of formal concept analysis (Carpineto and Romano 2004; Ganter and Wille 1999). Consider non-empty sets X and Y of objects and attributes, respectively, and a binary relation I (incidence relation) between X and Y. That is to say, an object $x \in X$ being in the relation I to an attribute $y \in Y$, which is denoted by $\langle x, y\rangle \in I$, indicates that x has y. The triplet $\langle X, Y, I\rangle$, called a formal context in terms of formal concept analysis, may be represented by a binary matrix such as the one in Table 1, in which the 1s (0s) represent that x_i has y_j (x_i does not have y_j).[3]

A formal concept in a given formal context $\langle X, Y, I\rangle$ is a pair $\langle A, B\rangle$ consisting of a set $A \subseteq X$ of objects (the so-called extent) and a set $B \subseteq Y$ of attributes (the so-called intent) satisfying

$$A^{\uparrow} = B \quad \text{and} \quad B^{\downarrow} = A, \tag{1}$$

where

$$\begin{aligned} A^{\uparrow} &= \{y \in Y \mid \text{for each } x \in A : \langle x, y\rangle \in I\} \text{ and} \\ B^{\downarrow} &= \{x \in X \mid \text{for each } y \in B : \langle x, y\rangle \in I\}. \end{aligned}$$

Condition (1) means that B is just the set of the attributes shared by all objects in A and A consists of the objects sharing all the attributes in B.

In Table 1, the pair consisting of $A = \{x_3, x_4\}$ and $B = \{y_2, y_3, y_4\}$ represents a formal concept. It may be understood as a category defined by a simultaneous presence of the binary features y_2, y_3 and y_4, or, put differently, a satisfaction of the three yes/no conditions represented by y_2, y_3 and y_4. Such a category includes the objects x_3 and x_4, but not, e.g. x_1 because x_1 does not possess y_4. Likewise, the pairs consisting of $A = \{x_1, x_2, x_3, x_4\}$ and $B = \{y_2, y_3\}$, and of $A = \{x_4\}$ and $B = \{y_1, y_2, y_3, y_4\}$, represent formal concepts. On the other hand, $A = \{x_4, x_5\}$ and $B = \{y_1, y_2\}$ do not form a concept because x_5 does not have y_2.

The set of all formal concepts in a given formal context $\langle X, Y, I\rangle$ is denoted by $\mathcal{B}(X, Y, I)$, i.e.

$$\mathcal{B}(X, Y, I) = \{\langle A, B\rangle \mid A^{\uparrow} = B \text{ and } B^{\downarrow} = A\},$$

and is called the concept lattice of $\langle X, Y, I\rangle$.

When equipped with a natural subconcept–superconcept hierarchy $\leq$, defined by

$$\langle A, B\rangle \leq \langle C, D\rangle \quad \text{if and only if} \quad A \subseteq C,$$

or, equivalently, if and only if $B \supseteq D$, the set $\mathcal{B}(X, Y, I)$ indeed becomes a partially ordered set, which in fact turns out to be a complete lattice, the structure of which is described by the so-called basic theorem of concept lattices (Ganter and Wille 1999).

2.2. Dutch data

The Dutch data (De Deyne et al. 2008) is a unique, rather extensive data which has been gathered by psychologists within a carefully designed study, in which hundreds of human respondents participated. The main theme of Dutch data is common language categories (concepts) and accompanying data which includes binary attributes (features) relevant to these categories, objects (exemplars) in these categories, and various psychologically relevant characteristics pertaining to these objects, attributes and categories.

For our purpose, we restrict to the part of the Dutch data concerning selected categories, the objects in these categories and relevant attributes. The data includes 16 linguistic categories, which consist of 10 natural-kind categories, 6 of which belong to the animal domain. In addition, it includes 6 categories of the artifact domain. Each category is represented by a set of objects (exemplars), such as a robin for the category “bird”. The exemplars were collected in an exemplar-listing process involving 527 participants in a previous study (Ruts et al. 2004). For each category, a number of exemplars, aimed at 30 per category, were selected from the listed exemplars, including typical and atypical ones, with the restriction that they are familiar to the vast majority of participants (Ruts et al. 2004).

We do not utilize the four non-animal natural kind categories as these are not part of the larger object-attribute data needed to conduct the intended experiments. The natural-kind categories we utilize are the 6 animal categories “amphibians”(includes 5 exemplars), [4] “birds” (30), “fish” (23), “insects” (26), “mammals” (30) and “reptiles” (22).[5] The 6 artifact categories are: “clothing” (29), “kitchen utensils” (33), “musical instruments” (27), “tools” (30), “vehicles” (30) and “weapons” (20).[6] For convenience, the exemplars of all the involved categories are described in Table A1 in Appendix 1.

The animal and the artifact categories comprise 129 and 166 exemplars, respectively, which are representative of these categories.[7] Coverage by the animal, as well as the artifact categories, is considerable (for instance, the animal categories cover a rather large part of the known animal domain).

The attributes (features) which describe the objects of the particular categories were obtained from 1003 respondents in two ways (we refer to De Deyne et al. (2008) for details): First, the respondents were required to list relevant attributes for a given category (these are called the category attributes). Second, they were asked to list relevant attributes for each object involved in the categories (exemplar attributes).

The numbers of objects in the two domains mentioned above and the numbers of attributes of the two kinds are displayed in Table 2. These numbers represent the four types of binary matrices used in our experiments, i.e. the 129×225 matrix of the animal domain with category attributes, 129×764 matrix of the animal domain with exemplar attributes, and the 166×301 and 166×1295 matrices of the artifact domain with category and exemplar attributes, respectively.

Table 2. Four types of binary matrices used in our experiments: objects in the animal/artifact domain with the respective category/exemplar attributes.

Domain	Objects	Category attributes	Exemplar attributes
Animal	129	225	764
Artifact	166	301	1295

The actual binary matrices corresponding to these four types are called the exemplar-by-feature applicability matrices by the authors of the Dutch data. They describe which objects have which attributes and represent a crucial component of the data for the questions we explore. Each of these matrices was completed separately by four respondents. Hence, there are four 129×225 binary matrices, corresponding to four respondents, and the same for the other three types of matrices.

In fact, the four respondents filling the 129×225 matrix with category attributes also filled the 166×301 matrix with category attributes, while the respondents filling the two matrices with exemplar attributes were eight distinct people (four per matrix). Note also that the matrices with the exemplar attributes were filled first and that the filled values for the exemplar attributes that were also among the category attributes[8] were copied to the 129×225 and 166×301 matrices with category attributes; the four respondents were then asked to complete the missing values in the two matrices with category attributes.[9]

The Dutch data also contains the corresponding four aggregated matrices, in which the values 0, 1, 2, 3 and 4 provide the number of respondents who agreed on that the respective object has the respective attribute. These matrices thus represent the strength of consensus among respondents as regards the presence of the attributes on the exemplars. There is hence one 129×225 consensus matrix with values equal to 0, 1, 2, 3 or 4, and the same for the remaining three types of matrices.

To obtain binary matrices from the consensus matrices, one naturally thresholds the matrix entries. For instance, from the 129×225 matrix with values in $\{0, 1, 2, 3, 4\}$ a binary matrix corresponding to the threshold "≥ 2" is obtained which contains 1 at row x and column y iff at least two respondents agreed that the object x has the attribute y (i.e. iff the value at row x and column y in the consensus matrix is ≥ 2). To illustrate, the matrix on the left (a made-up consensus matrix) gets transformed to the one on the right (corresponding binary matrix):

$$\begin{pmatrix} 0 & 2 & 3 & 1 \\ 1 & 4 & 4 & 0 \\ 0 & 3 & 4 & 3 \\ 2 & 2 & 2 & 4 \\ 2 & 1 & 0 & 1 \end{pmatrix} \xrightarrow{\text{“}\geq 2\text{”}} \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

This way, one obtains four thresholded binary matrices corresponding to the thresholds "≥ 1", "≥ 2", "≥ 3" and "$= 4$", for each of the four consensus matrices with dimensions described in Table 2, i.e. 16 binary matrices of considerable dimensions in total.

The original data contains some minor semantic and technical issues as regards a possible machine processing of the data. We hence modified the Dutch data, which is now available, along with a convenient Python wrapper, on GitHub (Belohlavek and Mikula 2023).

3. Experiments

3.1. The logic of our experiments

As described in Section 2.2, the Dutch data contains several binary matrices I each of which describes a set X of all exemplars of a given collection of categories using a set Y of binary attributes. As further explained above, these matrices are of four dimensions, namely 129×225 (animal domain with category attributes), 129×764 (animal domain with exemplar attributes), 166×301 (artifact domain with category attributes) and 166×1295 (artifact domain with exemplar attributes). For each of these dimensions, there are four binary matrices corresponding to four respondents who filled the matrices, and four thresholded binary matrices corresponding to the thresholds "≥ 1", "≥ 2", "≥ 3" and "$= 4$", which result from the consensus matrices. In total, we hence have 16 binary matrices describing the animal domain and 16 ones for the artifact domain: For each domain, $16 = 2$ (category or exemplar attributes) $\times 8$ (4 respondents' matrices plus 4 thresholded matrices).[10]

We may hence compute the corresponding 16 concept lattices $\mathcal{B}(X, Y, I)$ for the animal domain and the 16 concept lattices for the artifact domain. For each such concept lattice $\mathcal{B}(X, Y, I)$ and each given Dutch data category of the respective domain, which is represented by a subset C of the set X of all exemplars of the domain, we may then ask whether the category actually represents a formal concept in $\mathcal{B}(X, Y, I)$, i.e. whether

$$C = A \text{ for some formal concept } \langle A, B\rangle \in \mathcal{B}(X, Y, I). \tag{2}$$

Due to a basic property of the concept-forming operators (Ganter and Wille 1999), (2) holds true if and only if

$$C = (C^{\uparrow})^{\downarrow} \tag{3}$$

with $^{\uparrow}$ and $^{\downarrow}$ being the operators associated with the matrix I; see Section 2.1.

Since the categories, the exemplars, the binary features, as well as the binary exemplar-feature matrices have been obtained from a large number of human respondents via a carefully designed process, this kind of experiment may indicate whether the mathematical notion of a formal concept indeed represents a psychologically plausible model of human concepts. This is the principal question we address in our experiments. In addition, we also explore various related topics which naturally appear when exploring the principal question.

3.2. Closer look at relevant aspects of Dutch data

Before turning to the exploration outlined in the previous section, we examine additional characteristics and aspects of the data relevant to our principal question. These characteristics are not part of the Dutch data and may be of interest for other studies involving the data as well.

The first concerns the overlap between the collections of category and exemplar attributes for each domain. According to Section 2.2, the overlap affects the resulting binary matrices. As Table 3 shows, the overlap is considerable: More than 50 % of the category attributes appear among the exemplar attributes in both domains. This implies that because of the way the matrices have been gathered, more than half of the matrices with category

Table 3. Overlap of category and exemplar attributes for the animal and the artifact domain.

Domain	Objects	Category attributes	Exemplar attributes	Attribute overlap
Animal	129	225	764	129
Artifact	166	301	1295	176

Table 4. Density of matrices of individual respondents.

			Respondent			
Domain	Attribute type	Dimensions	1	2	3	4
Animal	Category	129×225	0.30	0.27	0.26	0.33
Artifact	Category	166×301	0.21	0.19	0.21	0.20
Animal	Exemplar	129×764	0.12	0.11	0.09	0.17
Artifact	Exemplar	166×1295	0.09	0.07	0.10	0.07

Table 5. Density of the thresholded matrices.

Domain	Attribute type	Dimensions	≥ 1	≥ 2	≥ 3	$= 4$
Animal	Category	129×225	0.43	0.32	0.25	0.17
Artifact	Category	166×301	0.34	0.23	0.15	0.09
Animal	Exemplar	129×764	0.23	0.13	0.08	0.04
Artifact	Exemplar	166×1295	0.16	0.09	0.05	0.02

attributes were copied from the corresponding matrices with exemplar attributes. As a consequence, the corresponding matrices with category and exemplar attributes are consistent regarding the presence or absence of the overlap attributes on the exemplars.

Another characteristic of possible interest is the density of the involved binary matrices, i.e. the proportion of the entries containing 1 among all entries in a given matrix. The densities of all the matrices are provided by Tables 4 and 5. While Table 5 only reflects the logical rule of decreasing densities as the threshold for consensus increases, Table 4 reveals a notable fact: For all matrices except for those for the animal domain with exemplar attributes, the densities corresponding to individual respondents are comparable, which indicates a reasonable consistency of respondents. For the matrix for animal domain and exemplar attributes, the density corresponding to respondent 4 is almost twice as high as the densities corresponding to the first three respondents. We hypothesize that this may be a result of a careless filling by the fourth respondent or of an excessively long contemplation of the respondent, which – as a rule – results in filling more 1s.

An interesting point with regard to the main question we explore is consistency among respondents as regards their judgment of whether a given attribute applies or does not apply to a given object. This information is provided by Tables 6, 7 and 8, which represent agreement among all the 6 pairs, all the 4 triples and of the whole quadruple of respondents, respectively. As an example, the column labelled "1, 2" in Table 6 represents the agreement of respondents 1 and 2: The values a, b, c and d are the numbers of entries defined as follows:

$$
\begin{aligned}
a &: \text{respondent 1 entered 1, respondent 2 entered 1,} \\
b &: \text{respondent 1 entered 1, respondent 2 entered 0,} \\
c &: \text{respondent 1 entered 0, respondent 2 entered 1,} \\
d &: \text{respondent 1 entered 0, respondent 2 entered 0.}
\end{aligned}
$$

Table 6. Agreement of pairs of respondents.

		1, 2	1, 3	1, 4	2, 3	2, 4	3, 4	mean
Animal Category	a	6343	5949	7272	5875	6666	6337	
	b	2411	2805	1482	2045	1254	1150	
	c	1577	1538	2387	1612	2993	3322	
	d	18694	18733	17884	19493	18112	18216	
129×225	SMC	0.86	0.85	0.87	0.87	0.85	0.85	0.86
Artifact Category	a	6754	6935	6787	6512	6533	6440	
	b	3900	3719	3867	2754	2733	4217	
	c	2512	3722	3101	4145	3355	3448	
	d	36800	35590	36211	36555	37345	35861	
166×301	SMC	0.87	0.85	0.86	0.86	0.88	0.85	0.86
Animal Exemplar	a	6604	5750	8368	6075	8081	6970	
	b	4989	5843	3225	4519	2513	2159	
	c	3990	3379	8402	3054	8688	9799	
	d	82974	83585	78562	84908	79274	79628	
129×764	SMC	0.91	0.91	0.88	0.92	0.89	0.88	0.9
Artifact Exemplar	a	9893	10842	9184	9875	8462	8643	
	b	9052	8103	9762	5440	6853	12140	
	c	5422	9941	4905	10908	5627	5446	
	d	190603	186084	191120	188747	194029	188742	
166×1295	SMC	0.93	0.92	0.93	0.92	0.94	0.92	0.93

Table 7. Agreement of triples of respondents.

		1, 2, 3	1, 2, 4	1, 3, 4	2, 3, 4	mean
Animal Category	a	5201	5808	5501	5426	
	d	17830	17165	17182	17411	
129×225	SMC	0.79	0.79	0.78	0.79	0.79
Artifact Category	a	5220	5360	5231	5085	
	d	34370	34872	33698	34555	
166×301	SMC	0.79	0.81	0.78	0.79	0.79
Animal Exemplar	a	4588	5842	5122	5326	
	d	81082	76812	77032	77864	
129×764	SMC	0.87	0.84	0.83	0.84	0.85
Artifact Exemplar	a	7061	6558	6504	6089	
	d	183476	187603	183319	185675	
166×1295	SMC	0.89	0.9	0.88	0.89	0.89

The table also contains degrees of agreement of the pairs of respondents expressed by the simple matching coefficient (SMC), which is defined as $\frac{a+d}{a+b+c+d}$, i.e. as the proportion of entries in which both respondents agree (Sokal and Michener 1958), as well as the mean value of SMC over all the 6 pairs of respondents. Tables 7 and 8 represent analogous information for the triples and the quadruple of respondents. In this case, we use a natural generalization of the SMC defined as $\frac{a+d}{n}$, where the values a and d denote the number of entries for which all the respondents of the triple or the quadruple entered 1 and 0, respectively, and n denotes the number of all entries of the binary matrix.

As one may observe, the agreement of the pairs, as well as the triples and the quadruple of respondents, is considerably high. This also justifies our consideration of the matrices

Table 8. Agreement of the quadruple of respondents.

		1, 2, 3, 4
Animal	a	4919
Category	d	16630
129 × 225	SMC	0.74
Artifact	a	4394
Category	d	32960
166 × 301	SMC	0.75
Animal	a	4318
Exemplar	d	75760
129 × 764	SMC	0.81
Artifact	a	4925
Exemplar	d	181451
166 × 1295	SMC	0.87

obtained from the consensus matrices by thresholds "$\geq i$" for $i = 1, 2, 3, 4$. In addition, notice that the somewhat abnormal density of the animal domain matrix with exemplar attributes of respondent 4 is apparent in Table 6: In all the columns "1, 4", "2, 4" and "3, 4", the values b are significantly smaller than the corresponding values c, indicating that respondent 4 entered considerably more 1s compared to respondents 1, 2 and 3.

3.3. Answers to main questions

We now present basic results regarding the question of whether the 11 human categories of the Dutch data turn out to be formal concepts in each of the available binary matrices, i.e. the thresholded consensus matrices and the individual respondent's matrices. The results are provided in Table 9 (thresholded consensus matrices) and Table 10 (individual respondents' matrices).

3.3.1. Main observations

3.3.1.1. Thresholded consensus matrices. We first turn to Table 9, where the corresponding binary matrices are derived from consensus among respondents and may hence be considered as representing a common view regarding the presence or absence of attributes. The table rows correspond to the inspected categories of the animal and the artifact domain (the first five and the next six categories, respectively). The columns represent the binary matrices for these two domains derived from the $\{0, 1, 2, 3, 4\}$-valued consensus matrices with the category attributes via the thresholds "≥ 1", "≥ 2", "≥ 3" and "$= 4$". The entry at a row corresponding to a category C and a column corresponding to a binary matrix I contains "yes" if C represents a formal concept in the corresponding concept lattice $\mathcal{B}(X, Y, I)$, i.e. if (3) holds, and "no" otherwise. Each table hence represents 44 tests (11 categories $\times 4$ matrices). In total, Table 9 represents 88 tests of whether a category is a formal concept in a reasonably elaborated description of exemplars by binary attributes, of which 40 are for the animal and 48 for the artifact domain.

Table 9. Are human categories formal concepts? Thresholded consensus matrices.

Attribute type	Domain	Category	≥ 1	≥ 2	≥ 3	$= 4$
Category	Animal	Bird	yes	yes	yes	yes
		Fish	yes	yes	yes	yes
		Insect	yes	yes	yes	yes
		Mammal	yes	yes	yes	no
		Reptile	yes	yes	yes	no
	Artifact	Clothing	yes	yes	yes	no
		Kitchen utensil	yes	yes	yes	no
		Musical instrument	yes	yes	yes	yes
		Tool	yes	yes	no	no
		Vehicle	no	no	no	no
		Weapon	no	no	no	no
Exemplar	Animal	Bird	yes	yes	yes	yes
		Fish	yes	yes	yes	yes
		Insect	yes	yes	yes	yes
		Mammal	yes	yes	yes	yes
		Reptile	yes	yes	yes	no
	Artifact	Clothing	yes	yes	yes	yes
		Kitchen utensil	yes	yes	yes	no
		Musical instrument	yes	yes	yes	yes
		Tool	yes	yes	no	no
		Vehicle	yes	no	no	no
		Weapon	no	no	no	no

A summary of the results of our tests is provided in the following table:

	Animal domain	Artifact domain	All
no. yes/no. tests	37/40	26/48	63/88

Let us mention a considerable sensitivity of the notion of formal concept to the description of objects by binary attributes, which is relevant in the assessment of these results. The absence of a single attribute on a given object may result in excluding the object from the extent of a given formal concept. Due to the large number of attributes and due to possible mistakes and flaws in the judgment of the respondents, both "false negatives" and "false positives" are likely to occur in the binary matrices involved in the Dutch data. In this perspective, the overall number of tests in which a category indeed turned out to represent a formal concept may be regarded as considerably high.

Intuitively, the matrices corresponding to thresholds "≥ 2" and "≥ 3" seem most natural since they represent a reasonable consensus of the respondents as regards the presence of attributes on exemplars. The matrix corresponding to "$= 4$" represents the maximal possible consensus, which arguably is not natural to be considered a common human view of the domain. Likewise, the matrix corresponding to "≥ 1" may be considered somewhat extreme since it displays a presence of an attribute whenever any of the respondents claims this presence (the matrix hence equals the union of the individual respondents' matrices).

For the sake of completeness and a possible further exploration, we include the list of all the formal concepts corresponding to all the entries of Table 9 in Tables A2 (category attributes) and A3 (exemplar attributes) in Appendices 2 and 3, respectively.[11] In particular, for each category represented by a set C of exemplars and each threshold "$\geq i$" of the

Table 10. Are human categories formal concepts? Individual respondents' matrices.

Attribute type	Domain	Category	Respondent 1	2	3	4
Category	Animal	Bird	yes	yes	yes	yes
		Fish	yes	yes	yes	yes
		Insect	yes	yes	yes	yes
		Mammal	yes	yes	yes	yes
		Reptile	yes	yes	yes	yes
	Artifact	Clothing	yes	yes	yes	no
		Kitchen utensil	no	yes	no	no
		Musical instrument	yes	yes	yes	yes
		Tool	no	yes	yes	no
		Vehicle	no	no	yes	no
		Weapon	no	no	no	no
Exemplar	Animal	Bird	yes	yes	yes	yes
		Fish	yes	yes	yes	yes
		Insect	yes	no	yes	yes
		Mammal	yes	yes	yes	yes
		Reptile	yes	yes	yes	yes
	Artifact	Clothing	yes	yes	yes	no
		Kitchen utensil	yes	yes	yes	no
		Musical instrument	yes	yes	yes	no
		Tool	yes	yes	yes	no
		Vehicle	yes	yes	yes	no
		Weapon	no	no	no	no

respective domain matrix, we list the formal concept $\langle C^{\uparrow\downarrow}, C^{\uparrow}\rangle$ generated by C by listing the objects of its extent and the attributes of its intent. We also include "yes" if the category forms a formal concept, i.e. $C = C$, and "no" otherwise. If the category does not form a formal concept, then $C \subset C^{\uparrow\downarrow}$, in which case we include the names of additional objects in $C^{\uparrow\downarrow} - C$ in italics.

Notice that for any given category C, the corresponding intent $C^{\uparrow_i}$ of the $\geq i$-thresholded matrix I gets larger as the threshold value i gets smaller. This is due to the following immediate consequence of the definition of the operator $\overset{\uparrow}{i}$ induced by the $\geq i$-thresholded
matrix I:

$$\text{If } i \leq j, \text{ then } C^{\uparrow_i} \supseteq C^{\uparrow_j} \text{ for any } C \subseteq X.$$

3.3.1.2. Animal domain and artifact domain. The numbers of positive and negative tests are apparently different for the animal and the artifact domain. In the animal domain, all tests except for three have positive results. The three negative results, however, occur in the matrix corresponding to the maximal possible respondents' consensus "= 4", which may be regarded as somewhat unnatural. The animal domain clearly supports the hypothesis that the notion of a formal concept provides a plausible model of human categories.

An important factor, which appears to contribute to the results, is the fact that all the categories in this domain are of the so-called natural kind.[12] Namely, the natural-kind categories tend to be more salient compared to artificial ones (Barsalou 1985), and "categorically distinct" in that "there cannot be a smooth transition from one kind to another"

(Bird and Tobin 2023). We thus contend that the natural-kind categories are better definable in terms of attributes generated by humans, as indicated by the animal-domain parts of Table 9.

In the artifact domain, the situation is different. Even when one disregards the matrix corresponding to the maximal consensus, i.e. the threshold "$= 4$", there remain categories, namely "tool", "vehicle" and "weapon", which do not form formal concepts. Consider "tool" first, which does not form a formal concept for thresholds "≥ 3" and "$= 4$" for either kind of attributes. For the threshold "≥ 3", the extent $C^{\uparrow\downarrow}$ includes (see Appendix 2) the set C of "tool" exemplars and, in addition, also some exemplars of "kitchen utensil" (10 for the category attributes and 19 for the exemplar attributes), such as a sieve, nutcracker, can opener and scissors. Clearly, the exemplars of "kitchen utensil" may rightly be regarded as exemplars of "tool" as well. Thus the reason for "tool" not being a formal concept in the "≥ 3"-thresholded consensus matrix is the natural overlap of "tool" and "kitchen utensil" that is not reflected in how these two categories are represented by exemplars in the Dutch data. A similar reason explains the situation with threshold "$= 4$", in which case the sets $C^{\uparrow\downarrow}$ contain 57 and 42 additional exemplars for the category and exemplar attributes, respectively.

The categories "vehicle" and "weapon", which fail to form formal concepts in 7 out of the 8 cases are somewhat more complicated. The presence of additional exemplars in $C^{\uparrow\downarrow}$ for the two sets C representing "vehicle" and "weapon" may partly be explained in a way similar to that for "tool". Intuitively, however, these categories, and the category "tool" as well, exemplify categories that are not as clearly separable as the natural-kind categories. A natural explanation of this intuition, which results in a failure to form a formal concept in most cases, is offered by Barsalou's research on goal-derived categories (Barsalou 1985). The goal-derived categories, of which "vehicle" and "weapon" are good examples, are formed with respect to the achievement of a certain goal. Such categories often include exemplars of several natural-kind categories, but not all exemplars of any such category. As a consequence, the exemplars of goal-oriented categories may be more difficult to define by binary features listed as exemplar or category attributes by respondents exposed to the exemplar and category names. Indeed, except for "vehicle" and the "≥ 1" threshold, the sets of binary attributes shared by the exemplars of the two categories, "vehicle" and "weapon", do not include attributes that convincingly represent the goals of the respective categories; cf. description of these categories in Appendix 2.

3.3.1.3. Individual respondents' matrices. In addition to the tests with the thresholded consensus matrices described above, we also provide in Table 10 results of the tests with the binary matrices filled by the individual respondents. These experiments test whether a category forms a formal concept with respect to individual person's knowledge rather than consensus knowledge. A summary of the results is the following:

	Animal domain	Artifact domain	All
No. yes/no. tests	40/40	26/48	66/88

In the animal domain, all tested categories turn out to form formal concepts irrespective of whether category attributes or exemplars attributes were used for the description of

Table 11. Number of formal concepts in the thresholded consensus matrices $I_{\geq i}$.

Domain	Attribute type	≥ 1	≥ 2	≥ 3	$= 4$
Animal	Category	11 097 215	1 669 900	156 632	28 212
Animal	Exemplar	671 126 463	13 258 687	236 387	21 148
Artifact	Category	12 853 601	3 243 995	725 063	54 292
Artifact	Exemplar	1 379 165 960	15 146 201	1 051 635	25 242

the objects. Note that this also is true for respondent 4 for the animal domain with exemplar attributes whose matrix contains considerably more 1s than those of the other three respondents; see Section 3.2. A possible explanation of this remarkable fact is that in spite of the higher density, respondent's 4 matrix contains a consistent view of the exemplars and features in which all the categories still are definable by the features.

The ratio of 26/48 in the artifact domain deserves a deeper look. First, we see a notable number of negative results for the category "weapon" and also for "vehicle", similarly as in the case of the threshold consensus matrices above. The reason for this is arguably similar to the one in the consensus case, namely the somewhat peculiar nature of these two categories (see above).

Notice also the results for respondents 4 in Table 10 containing "no" almost everywhere. Recall that these respondents are two different persons, which we now denote r4c and r4e, because they filled the matrices with category and exemplar attributes, respectively; due to how the matrices we filled, almost half of the matrix of r4c (the part corresponding to the overlap attributes) was copied from the matrix of r4e. We hypothesize that while the matrix of r4e, for whatever conceivable reason, simply does not render the artifact categories as formal concepts, the main reason for the "no" answers for the matrix of r4c is the copying of the values of overlapping attributes from the matrix of r4e. Namely, we observed that when these values were copied from the matrices of respondents 1, 2 and 3, the majority of the answers are positive.

3.3.2. Further observations

Let us now present additional observations on matters related to the main question explored in the previous section. The first concerns the following pattern present in Table 9. In both tables, it holds true that all the entries right to any "no"-entry also contain "no". That is, if a category fails to be a formal concept in a consensus matrix for threshold "$\geq i$", then it also fails to be a formal concept for all higher thresholds, i.e. for "$\geq j$" with $j > i$ (equivalently, if a category forms a formal concept for "$\geq i$ ", then it forms a formal concept for all lower thresholds). This pattern, however, is but a result of coincidence in the Dutch data, as one can easily construct data with matrices $I_{\geq i} > I_{\geq j}$ for $i < j$ and find a set C of objects which is an extent w.r.t. $I_{\geq j}$, but not w.r.t. $I_{\geq i}$.

Another observation concerns the size of the concept lattices $\mathcal{B}(I_{\geq i})$ of the thresholded consensus matrices $I_{\geq}$. Namely, it turns out that for both the animal and the artifact domain and for both category and exemplar attributes, the number of formal concepts gets smaller as the consensus threshold increases. That is,

$$i < j \text{ implies } |\mathcal{B}(I_{\geq i})| > |\mathcal{B}(I_{\geq j})|,$$

as is apparent from Table 11.

Table 12. No. cases in which a category form a formal concept for the thresholded consensus matrices in Table 9.

	cat = yes *exe* = yes	*cat* = yes *exe* = no	*cat* = no *exe* = yes	*cat* = no *exe* = no
Animal	18	0	1	1
Artifact	12	0	2	10
Both	30	0	3	11

Table 13. No. cases in which a category forms a formal concept for the individual respondents' matrices in Table 10.

	cat = yes *exe* = yes	*cat* = yes *exe* = no	*cat* = no *exe* = yes	*cat* = no *exe* = no
Animal	20	0	0	0
Artifact	10	1	5	8
Both	30	1	5	8

One might hence be tempted to conclude that, e.g. the formal concepts of $\mathcal{B}(I_{\geq 3})$ are included in $\mathcal{B}(I_{\geq 2})$. Such a conclusion would be false. In fact, it appears that even though $|\mathcal{B}(I_{\geq i})| > |\mathcal{B}(I_{\geq j})|$, only a very small fraction of the extents of formal concepts in $\mathcal{B}(I_{\geq j})$ actually are also extents of formal concepts in $\mathcal{B}(I_{\geq i})$.

Note also that one needs to be careful in drawing seemingly intuitive but, in fact, wrong conclusions about the relationships between the tables regarding the thresholded consensus matrices and the matrices corresponding to the individual responses. For example, it is not true that if a category forms a formal concept with regard to all four matrices of the individual respondents, then it forms a formal concept in the matrix corresponding to the maximum consensus, i.e. to the "= 4"-thresholded matrix, nor vice versa: The categories "reptile" and "clothing" in the matrices with category attributes serve as counterexamples.

Our next observation concerns the relationship between category and exemplar attributes, as regards their capability to define a category. In particular, we consider the confidence of the rules[13]

$cat \to exe$: if a category is a formal concept for *cat*egory attributes, it also is formal concept for *exe*mplar attributes,

and

$exe \to cat$: if a category is a formal concept for *exe*mplar attributes, it also is formal concept for *cat*egory attributes.

The relevant numbers of cases in which the inspected categories form formal concepts are provided by Table 12 (thresholded consensus matrices) and Table 13 (individual respondents' matrices).

For instance, the numbers 18 and 0 in the "animal" line of Table 12 indicate that in 18 of the 20 entries of the animal-domain part of Table 9, the value is "yes" for the table with category attributes and "yes" for the table with exemplar attributes; and, moreover, that in 0 of

Table 14. Confidence of rules $cat \rightarrow exe$ and $exe \rightarrow cat$ for data in Table 12 (left) and Table 13 (right).

	Animal	Artifact	Both		Animal	Artifact	Both
$cat \rightarrow exe$	1	1	1	$cat \rightarrow exe$	1	0.91	0.97
$exe \rightarrow cat$	0.95	0.86	0.91	$exe \rightarrow cat$	1	0.67	0.86

these 20 entries the respective values are "yes" for category attributes and "no" for exemplar attributes. The confidence values of both rules, $cat \rightarrow exe$ and $exe \rightarrow cat$, are provided in Table 14; the values on the left and on the right correspond to the thresholded consensus matrices and the individual respondents' matrices, respectively. The confidence values of $cat \rightarrow exe$ may be interpreted as confirming the intuition that the exemplar attributes have a better distinctive ability compared to the category ones: With only a few exceptions, if a category is definable using the category attributes, it also is definable using the exemplar attributes.

4. Conclusions and future research

4.1. Basic findings

Our primary goal is to explore the psychological plausibility of the notion of a formal concept using the now available but little-exploited high-quality psychological data referred to as the Dutch data. In a broader sense, we intend to bring attention to the psychological relevance of the formal notions utilized in data mining, machine learning, and artificial intelligence inspired by the human mind.

We assess our primary question by asking whether the human categories involved in the Dutch data form formal concepts in the several binary matrices provided in the data. In brief, the matrices represent binary data gathered from individual respondents and data representing a consensus among the respondents. The matrices concern two large domains, namely the animal domain and the artifact domain. They are available with two large collections of attributes: Category attributes and exemplar attributes.

It turns out that our main question of whether a given human category forms a formal concept has an affirmative answer for most of the available human categories and binary matrices. This supports the view that the notion of a formal concept indeed provides a psychologically plausible model of human categories. In addition to the main question, we examine various aspects of the Dutch data, such as the overlap of the category and exemplar attributes, the densities of the binary exemplar-feature matrices, and the consistency of data provided by the individual respondents, which provides a relevant context for our explorations and themselves are of psychological interest.

4.2. Future research topics

Our explorations shall lead to further research and a renewed interest in the classical view of concepts, both from the viewpoint of the psychology of concepts and from the perspective of utilizing formal models of concepts in applications. The classical view has been the prevailing approach since Aristotle. It dominated the psychological approaches to concepts before the 1970s when it came to be questioned, given the psychological studies by Eleanor

Rosch (Rosch 1978; Rosch and Mervis 1975; Rosch et al. 1976); see Murphy (2002) for a thorough exposition.

Recall that according to the classical view, a concept is determined by a collection of binary attributes with a defining role: The concept applies to an object if and only if the object has all these defining attributes. Rosch's findings brought to attention what is nowadays referred to as the graded structure of concepts, which has since been regarded as empirical evidence against the classical view. In brief, Rosch's experiments revealed that various phenomena, such as membership in a human concept (category), are a matter of degree rather than bivalent (yes–no), as the classical view assumes. In addition to empirical evidence, which is considered the main argument against the classical view, there is also the following in-principle argument against it, which is attributed to Wittgenstein (1953). Even though a majority of concepts seems to be defined by a collection of binary attributes, i.e. the necessary and sufficient conditions, for most concepts it appears to be impossible actually to specify the defining attributes. Whenever one suggests a collection of attributes as a definition of a given concept, there seems to pop up an object subsumed by the concept that does not satisfy the definition or, vice versa, an object meeting the definition is not subsumed by the concept. The above arguments led to dismissing the classical view as a viable approach within the psychology of concepts.

In our view, an overall dismissal of the classical view is inappropriate. On the one hand, the classical view is currently justly regarded as not accounting properly for several phenomena considered within the psychological research. In the end, the realm of human concepts is highly complex, and one may hardly hope that any given formal model of concepts copes with all the peculiarities involved to complete satisfaction. On the other hand, however, the classical view provides a rather appealing model of human concepts of considerable pragmatic value:

- Numerous publications in the psychology of concepts that appeared even a long time after the 1970s dismissal of the classical view still employ models in which human categories are represented by classical sets of exemplars further described by binary attributes.
- Binary attributes may naturally explain several graded phenomena accompanying human concepts. A case in point is the phenomenon of typicality of exemplars in human categories; see the above works of Rosch, the references in Murphy (2002) and our recent study (Belohlavek and Mikula 2022).
- The classical view provides at least a reasonable approximation of human concepts. It is useful in data management applications, particularly in analyzing data. Moreover, it may provide a useful, albeit simplified, formal model for psychological explorations, amenable to quantitative considerations which may help understand various psychological phenomena. With appropriate further developments, e.g. extensions and modifications, such as including graded (fuzzy) attributes instead of relying on only binary attributes, it may obtain a better psychological plausibility; see, e.g. Belohlavek and Klir (2011).

We, therefore, believe that further explorations in the classical view of concepts are needed that include an interaction of the psychological viewpoint and the formal (i.e.

logico-mathematical) viewpoint. Below we describe some particular problems in which we started to take the first steps.

A question that naturally offers itself is: If a human category forms a formal concept $\langle A, B\rangle$ with respect to a given exemplar-feature matrix, does the set of attributes B provide a natural definition of the category? Note first that the scenario of our experiments differs from the one commonly employed when reasoning about the classical view: The set of binary attributes defining a given category is not supplied by a human; instead, the defining attributes are sought in the exemplar-feature matrix.

An inspection of Appendix 2 reveals that the answer is affirmative for some categories and binary matrices with relatively small sets B, such as "bird" or "fish" in the matrices with the category as well as exemplar attributes corresponding to "≥ 3" or "$= 4$" or "clothing" for "≥ 3". For most categories, however, the corresponding set B of attributes may not be regarded as a natural definition of the category, the primary reason being the excessive number of attributes in B; see, e.g. "bird" for both types of matrices corresponding to "≥ 1". In our view, this is generally caused by the fact that some of the attributes in B are naturally considered more critical than others, and hence play a more important role for defining a category in a natural way.

One aspect in these considerations relates to the logical entailment of attributes. The possibly large set B may contain a smaller subset $B^* \subseteq B$ that entails every attribute in B in that every exemplar with all the attributes in the smaller B^* also has every attribute in B. Arguably, the attributes in B^* may be considered more essential for the given category than the entailed attributes in B that do not belong to B^*. For instance, for the category "bird" and the matrix with the category attributes corresponding to "≥ 2", the set B of the corresponding formal concept $\langle A, B\rangle$ consists of 33 attributes but includes a three-attribute set B^*, namely,

$$B^* = \{\text{has a bill},\ \text{has feathers},\ \text{has two paws}\},$$

that entails all the remaining 30 attributes in B. Arguably, B^* may be regarded as a reasonable definition of the category "bird". As one may see, the same set may work as a definition of "bird" with respect to the exemplar attributes.

It is a matter of an immediate observation that B^* entails B if and only if B^* generates B in that $(B^*)^{\downarrow\uparrow} = B$. When looking for natural definitions, one hence seems compelled to look for small generators B^*. We observed that it might not be best to look for generators with the smallest possible size or for generators that are minimal with respect to set inclusion. Namely, each of the three singleton subsets of the above three-attribute set B^* itself, i.e.

$$\{\text{has a bill}\}, \{\text{has feathers}\} \text{ and } \{\text{has two paws}\},$$

is a generator. But while B^* may be regarded as a reasonable definition of "bird", none of the singletons can. The principles according to which some attributes may be dropped require further study.

The preceding question of which attributes may be removed from an intent of a formal concept connects to a more general question of what makes a set B of attributes a good definition of a given category. In exploring this question, it seems that one shall proceed unconfined by the basic imperative of the classical view, namely, that an exemplar is a member of the category if and only if it satisfies all the attributes in B. Such a view allows no exceptions. Instead, one might consider more relaxed conditions which allow missing attributes for certain objects, such as "can fly" for a penguin in the category "bird". These considerations go beyond the classical view of concepts and may be regarded as its extension worth further exploration.

Note

1. See also the numerous foundational contributions and applications described in the papers of the three dedicated conferences, the ICFCA (Int. Conf. Formal Concept Analysis), the CLA (Concept Lattices and Their Applications), and ICCS (Int. Conf. Formal Concept Analysis).
2. Two exceptions we are aware of are the recent studies by Belohlavek and Trnecka (2020a, 2020b) and by Belohlavek and Mikula (2022).
3. For convenience and by a slight abuse of notation, we also say "binary matrix I" instead of "relation I".
4. As the category "amphibians" only contains 5 exemplars, which are all included in the category "reptiles", we omit it in most of our considerations below; see De Deyne et al. (2008) for reasons to include the exemplars of "amphibians" in "reptiles".
5. The binary matrices, which we describe below and use in our experiments, contain only 20 exemplars of the category "reptiles" because the respondents who were to fill in these matrices turned out not to be familiar with two exemplars, komodo and iguanodon (De Deyne et al. 2008).
6. Here, we use the plural in category names, as the authors do (De Deyne et al. 2008); below, we use singular, i.e. "bird" rather than "birds" to be consistent with our previous writings.
7. In addition to the 5 amphibians included in reptiles and two omitted exemplars of reptiles as mentioned above, three exemplars of the artifact categories are included in two distinct categories. There are no other overlaps of the categories.
8. Namely, the exemplar and category attributes overlap; see below.
9. The data also contains much smaller matrices corresponding to the individual categories. For every category, there are two matrices describing all exemplars of the category using binary attributes, one using the category attributes and the other using the exemplar attributes associated with the given category. These matrices are not interesting for the problem we examine since each such matrix describes just one category.
10. Recall that there were 12 respondents in total filling in all the 32 matrices: Four respondents, each filling a matrix describing the animal domain by category attributes, the same four respondents for the matrices describing the artifact domain by category attributes, additional four respondents for the matrices describing the animal domain by exemplar attributes, and yet other four respondents for the matrices describing the artifact domain by exemplar attributes. In what follows, we refer to the four respondents corresponding to each given matrix type as respondents 1, 2, 3, and 4, respectively.
11. We include the list, particularly for a possible exploration by the psychologists.
12. The term "natural kind" in relation to categories appears in psychology and philosophy. The adjective "natural" does not refer to the naturalness of the entities being grouped by a given category. Rather it refers to the naturalness of the grouping represented by the category in that it "that reflects the structure of the natural world rather than the interests and actions of human beings" (Bird and Tobin 2023). The notion of a natural kind is not clear cut, though, and the terminology is even ambiguous in the psychology of concepts. According to Murphy (2002, n. 4, p. 500), for instance,

    > [t]he term *natural categories* refers to categories that people naturally and normally use in everyday life – not to categories of nature. They are to be contrasted with *artificial categories*, which are made up by experimenters to test their theories. So, furniture and guns would be considered natural categories, because these are categories people use in everyday life. Unfortunately, there is another term called *natural kinds*, which does refer in part to categories of nature."

13. A confidence of a rule $\varphi \to \psi$ is the defined as $\frac{m}{n}$, where m is the number of cases in which both φ and ψ are true, and n is the number of cases in which φ is true.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Supported by grants IGA_PrF_2022_018, IGA_PrF_2023_026, and partly by the 2024 IGA grant of the Palacký University Olomouc.

ORCID

Radim Belohlavek http://orcid.org/0000-0003-4924-3233
Tomas Mikula http://orcid.org/0000-0002-7514-8523

References

Arnauld, Antoine, and Pierre Nicole. 1962. *La Logique, ou L'art de Penser*. 1st ed. Paris: Jean Guignart, Charles Savreux, & Jean de Lavnay. English ed. 1996. *Logic or the Art of Thinking*. Translated by J. V. Buroker. Cambridge: Cambridge University Press.

Barsalou, Lawrence W. 1985. "Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure in Categories." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11 (4): 629–654.

Belohlavek, Radim, and George J. Klir, eds. 2011. *Concepts and Fuzzy Logic*. Cambridge, MA: The MIT Press.

Belohlavek, Radim, and Tomas Mikula. 2022. "Typicality: A Formal Concept Analysis Account." *International Journal of Approximate Reasoning* 142: 349–369. https://doi.org/10.1016/j.ijar.2021.12.001.

Belohlavek, Radim, and Tomas Mikula. 2023. "Dutch Concepts" (dataset). V. 0.1.2. GitHub repository. https://github.com/mikulatomas/dutch-concepts.

Belohlavek, Radim, and Martin Trnecka. 2020a. "Basic Level of Concepts in Formal Concept Analysis 1: Formalization and Utilization." *International Journal of General Systems* 49 (7): 689–706. https://doi.org/10.1080/03081079.2020.1837794.

Belohlavek, Radim, and Martin Trnecka. 2020b. "Basic Level of Concepts and Formal Concept Analysis 2: Examination of Existing Basic Level Metrics." *International Journal of General Systems* 49 (7): 707–723. https://doi.org/10.1080/03081079.2020.1837795.

Bird, Alexander, and Emma Tobin. 2023. "Natural Kinds." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman (Spring 2023 Edition) https://plato.stanford.edu/archives/spr2023/entries/natural-kinds/.

Carpineto, Claudio, and Giovanni Romano. 2004. *Concept Data Analysis: Theory and Applications*. Chichester, England, Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/0470011297.

De Deyne, Simon, Steven Verheyen, Eef Ameel, Wolf Vanpaemel, Matthew J. Dry, Wouter Voorspoels, and Gert Storms. 2008. "Exemplar by Feature Applicability Matrices and Other Dutch Normative Data for Semantic Concepts." *Behavior Research Methods* 40 (4): 1030–1048. http://dx.doi.org/10.3758/BRM.40.4.1030.

Ganter, Bernhard, Sergei Obiedkov, Sebastian Rudolph, and Gerd Stumme. 2016. *Conceptual Exploration*. Heidelberg: Springer.

Ganter, Bernhard, Gerd Stumme, and Rudolf Wille, eds. 2005. *Formal Concept Analysis: Foundations and Applications*. Vol. 3626. Berlin: Springer.

Ganter, Bernhard, and Rudolf Wille. 1999. *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer. http://dx.doi.org/10.1007/978-3-642-59830-2.

Machery, Edouard. 2007. "100 Years of Psychology of Concepts: The Theoretical Notion of Concept and Its Operationalization." *Studies in History and Philosophy of Biological and Biomedical Sciences* 38 (1): 63–84. https://doi.org/10.1016/j.shpsc.2006.12.005.

Murphy, Gregory. 2002. *The Big Book of Concepts*. Cambridge, MA: MIT Press. http://dx.doi.org/10.7551/mitpress/1602.001.0001.

Rosch, Eleanor. 1978. "Principles of categorization." In *Cognition and Categorization*, edited by Eleanor Rosch and Barbara B. Lloyd, 27–48. Hillsdale, NJ: Erlbaum.

Rosch, Eleanor, and Carolyn B. Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7 (4): 573–605. https://doi.org/10.1016/0010-0285(75)90024-9.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem. 1976. "Basic Objects in Natural Categories." *Cognitive Psychology* 8 (3): 382–439. https://doi.org/10.1016/0010-0285(76)90013-X.

Ruts, Wim, Simon De Deyne, Eef Ameel, Wolf Vanpaemel, Timothy Verbeemen, and Gert Storms. 2004. "Dutch Norm Data for 13 Semantic Categories and 338 Exemplars." *Behavior Research Methods, Instruments, & Computers* 36 (3): 506–515. https://doi.org/10.3758/BF03195597.

Smith, Edward E., and Douglas L. Medin. 1981. *Categories and Concepts*. Cambridge, MA: Harvard University Press. http://dx.doi.org/10.7551/mitpress/1602.001.0001.

Sokal, Robert Reuven, and Charles Duncan Michener. 1958. "A Statistical Method for Evaluating Systematic Relationships." *The University of Kansas Science Bulletin* 38 (Part 2(22)): 1409–1438.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Appendices

Appendix 1. Categories and their exemplars

Table A1. The categories involved in the Dutch data.

Category	Count	Exemplars
Bird	30	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker
Fish	23	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale
Insect	26	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm
Mammal	30	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra
Reptile	20	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper
Clothing	29	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit
Kitchen utensil	33	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok
Musical instrument	27	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin
Tool	30	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench
Vehicle	30	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin
Weapon	20	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip

Appendix 2. Formal concepts of the exemplar-by-feature binary matrices

Table A2. Categories and formal concepts of the animal and artifact domains with category attributes.

Category	Threshold	Is formal concept?	Extent	Intent
Bird	≥ 1	Yes	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker	belongs to nature, breathes, breathes through lungs, builds nests, can become ill, can walk, carries over diseases, comes in different sizes, contains proteins, descends from reptiles, dies, does not smell well, doesn't resemble a human being, eats, eats insects, eats small animals, eats worms, exists for ages, flies or crawls, has a beak, has a bill, has a central nervous system, has a leathery skin, has a mouth, has a tail, has air sacs, has an aerodynamic body, has brains, has eyes, has feathers, has legs (poten), has lungs, has no eyelids, has no hairs, has sharp claws (nagels), has skin, has two eyes, has two paws, has two wings, has wings, is a bird, is a carnivore, is a collective noun, is a living being, is a vertebrate, is able to reproduce, is an animal, is eaten by other animals, is fascinating, is not a mammal, is not dangerous, is not yet independent at birth, is soft, is useful, is warm-blooded, its vascular system is less developed than that of mammals, lays eggs, lives in the wild, lives on land, makes a lot of noise, makes a sound, sometimes smells, there are lots of these, there are many kinds of it
Bird	≥ 2	Yes	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker	belongs to nature, breathes, breathes through lungs, builds nests, can become ill, can walk, comes in different sizes, contains proteins, descends from reptiles, dies, doesn't resemble a human being, eats, has a beak, has a bill, has a central nervous system, has a mouth, has air sacs, has brains, has eyes, has feathers, has legs (poten), has lungs, has no hairs, has skin, has two eyes, has two paws, has two wings, has wings, is a bird, is a living being, is a vertebrate, is able to reproduce, is an animal, is eaten by other animals, is not a mammal, is not yet independent at birth, is useful, is warm-blooded, lays eggs, lives on land, makes a lot of noise, makes a sound, sometimes smells, there are lots of these
Bird	≥ 3	Yes	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker,	belongs to nature, breathes, breathes through lungs, can become ill, contains proteins, descends from reptiles, dies, doesn't resemble a human being, eats, has a beak, has a bill, has a central nervous system, has brains, has eyes, has legs (poten), has lungs, has no hairs, has skin, has two eyes, has two paws, has wings, is a bird, is a living being, is a vertebrate, is able to reproduce, is an animal, is not a mammal, is not yet independent at birth, is useful, is warm-blooded, lays eggs, makes a sound
Bird	≥ 4	Yes	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker	breathes, can become ill, contains proteins, dies, doesn't resemble a human being, eats, has a bill, has a central nervous system, has brains, has eyes, has legs (poten), has lungs, has skin, has two eyes, has two paws, has wings, is a bird, is a living being, is a vertebrate, is able to reproduce, is an animal, is not a mammal

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Fish	≥ 1	Yes	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale	belongs to nature, breathes, breathes under water, can become ill, can swim, can't fly, comes in different sizes, contains proteins, dies, doesn't have wings, doesn't live on land, doesn't resemble a human being, eats, eats small animals, eats worms, exists for ages, has a central nervous system, has a mouth, has a tail, has an aerodynamic body, has brains, has eyes, has fins, has gills, has no eyelids, has no hairs, has no paws, has short paws, has teeth, has two eyes, herds, is a carnivore, is a collective noun, is a living being, is a vertebrate, is able to reproduce, is an animal, is cold-blooded, is eaten by other animals, is fascinating, is fast, is slippery, is smooth, is sometimes eaten by man, is useful, lives in Africa, lives in the sea, lives in the wild, lives in warm countries, lives in water, lives nearby the water, makes a lot of noise, shines, sometimes smells, there are lots of these, there are many kinds of it
Fish	≥ 2	Yes	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale	belongs to nature, breathes, breathes under water, can become ill, can swim, can't fly, comes in different sizes, contains proteins, dies, doesn't have wings, doesn't live on land, doesn't resemble a human being, eats, eats small animals, has a central nervous system, has a mouth, has an aerodynamic body, has brains, has eyes, has fins, has gills, has no hairs, has no paws, has teeth, has two eyes, is a living being, is able to reproduce, is an animal, is eaten by other animals, is slippery, is smooth, is useful, lives in the sea, lives in the wild, lives in warm countries, lives in water, sometimes smells
Fish	≥ 3	Yes	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale	belongs to nature, breathes, can become ill, can swim, can't fly, contains proteins, dies, doesn't have wings, doesn't live on land, doesn't resemble a human being, eats, has a central nervous system, has brains, has eyes, has gills, has no hairs, has no paws, has two eyes, is a living being, is able to reproduce, is an animal, is slippery, is smooth, is useful, lives in the wild, lives in water
Fish	≥ 4	Yes	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale	can become ill, can swim, can't fly, dies, doesn't have wings, doesn't resemble a human being, eats, has a central nervous system, has eyes, has no hairs, has two eyes, is a living being, is able to reproduce, is an animal, is slippery, is smooth, lives in water

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Insect	≥ 1	Yes	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm	Belongs to nature, breathes, builds nests, can become ill, comes in different sizes, contains proteins, dies, does not taste well, doesn't have a central nervous system, doesn't resemble a human being, eats, exists for ages, has a central nervous system, has a mouth, has an exoskeleton, has brains, has eyes, has feelers, has no eyelids, has no hairs, has short paws, has teeth, has two eyes, herds, invertebrate, is a collective noun, is a living being, is able to reproduce, is an animal, is athrodopal, is eaten by an insectivore, is eaten by birds, is eaten by other animals, is exterminated, is fascinating, is found in the garden, is killed by man, is more often encountered during summer, is not a mammal, is not very popular among man, is small (klein), is smaller than a meter, is useful, its vascular system is less developed than that of mammals, lays eggs, lives in Africa, lives in the wild, lives in warm countries, lives on land, sometimes smells, there are lots of these, there are many kinds of it, uses the sun to keep body temperature stable
Insect	≥ 2	Yes	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm	belongs to nature, breathes, can become ill, comes in different sizes, contains proteins, dies, does not taste well, doesn't have a central nervous system, doesn't resemble a human being, eats, has a mouth, has brains, has eyes, has feelers, has no eyelids, has no hairs, has short paws, has two eyes, invertebrate, is a living being, is able to reproduce, is an animal, is eaten by other animals, is found in the garden, is not a mammal, is small (klein), is smaller than a meter, is useful, its vascular system is less developed than that of mammals, lays eggs, lives in Africa, lives in the wild, lives in warm countries, lives on land, there are lots of these, uses the sun to keep body temperature stable
Insect	≥ 3	Yes	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm	belongs to nature, breathes, can become ill, contains proteins, dies, does not taste well, doesn't have a central nervous system, doesn't resemble a human being, eats, has brains, has no eyelids, has no hairs, invertebrate, is a living being, is able to reproduce, is an animal, is not a mammal, is small (klein), is smaller than a meter, is useful, its vascular system is less developed than that of mammals, lays eggs, lives in the wild, lives on land, uses the sun to keep body temperature stable
Insect	≥ 4	Yes	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm	breathes, dies, doesn't resemble a human being, has no eyelids, invertebrate, is a living being, is an animal, is not a mammal, is small (klein), is smaller than a meter, its vascular system is less developed than that of mammals, lives on land

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Mammal	≥ 1	Yes	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra	belongs to nature, breastfeeds its babies, breathes, breathes through lungs, can become ill, can bite, can walk, comes in different sizes, contains proteins, descends from reptiles, dies, does not lay eggs, does not smell well, eats, exists for ages, has a central nervous system, has a fur (vacht), has a leathery skin, has a mouth, has a tail, has brains, has eyes, has four paws, has legs (poten), has lungs, has nipples, has skin, has teeth, has two eyes, is a carnivore, is a collective noun, is a living being, is a vertebrate, is able to reproduce, is an animal, is eaten by other animals, is fascinating, is hairy, is not yet independent at birth, is useful, is warm-blooded, its babies are born alive, lives in Africa, lives on land, makes a lot of noise, makes a sound, mammal, sometimes smells, there are lots of these, there are many kinds of it, thick skin
Mammal	≥ 2	Yes	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra	belongs to nature, breastfeeds its babies, breathes, breathes through lungs, can become ill, comes in different sizes, contains proteins, dies, does not lay eggs, eats, has a central nervous system, has a mouth, has a tail, has brains, has eyes, has legs (poten), has lungs, has nipples, has skin, has teeth, has two eyes, is a living being, is a vertebrate, is able to reproduce, is an animal, is eaten by other animals, is not yet independent at birth, is useful, is warm-blooded, its babies are born alive, lives on land, makes a sound, mammal, sometimes smells
Mammal	≥ 3	Yes	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra	belongs to nature, breastfeeds its babies, breathes, breathes through lungs, can become ill, contains proteins, dies, does not lay eggs, eats, has a central nervous system, has a mouth, has brains, has eyes, has legs (poten), has lungs, has nipples, has skin, has teeth, has two eyes, is a living being, is a vertebrate, is able to reproduce, is an animal, is not yet independent at birth, is useful, is warm-blooded, its babies are born alive, lives on land, mammal
Mammal	≥ 4	No	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra, *dolphin*, *whale*, *orca*	breathes, breathes through lungs, can become ill, contains proteins, dies, eats, has a central nervous system, has brains, has eyes, has lungs, has skin, has two eyes, is a living being, is a vertebrate, is able to reproduce, is an animal, mammal

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Reptile	≥ 1	Yes	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper	belongs to nature, breathes, breathes through lungs, builds nests, can become ill, can swim, can't fly, comes in different sizes, contains proteins, dark colour, descends from reptiles, dies, does not taste well, doesn't have wings, doesn't resemble a human being, eats, eats insects, eats plants, eats small animals, eats worms, exists for ages, has a central nervous system, has a mouth, has a tail, has brains, has eyes, has lungs, has no eyelids, has no hairs, has scales, has short paws, has skin, has teeth, has two eyes, is a carnivore, is a collective noun, is a hunter, is a living being, is a vertebrate, is able to reproduce, is an animal, is cold-blooded, is eaten by other animals, is fascinating, is not a mammal, is smaller than a meter, is ugly, is useful, its vascular system is less developed than that of mammals, lays eggs, lives in Africa, lives in the wild, lives in the zoo, lives in warm countries, lives nearby the water, lives on land, makes a lot of noise, sometimes smells, there are lots of these, there are many kinds of it, uses the sun to keep body temperature stable
Reptile	≥ 2	Yes	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper	belongs to nature, breathes, breathes through lungs, can become ill, can't fly, comes in different sizes, contains proteins, dies, does not taste well, doesn't have wings, doesn't resemble a human being, eats, eats insects, eats small animals, has a central nervous system, has a mouth, has brains, has eyes, has lungs, has no eyelids, has no hairs, has two eyes, is a living being, is a vertebrate, is able to reproduce, is an animal, is cold-blooded, is not a mammal, is useful, its vascular system is less developed than that of mammals, lays eggs, lives in Africa, lives in the wild, lives in warm countries, sometimes smells, uses the sun to keep body temperature stable
Reptile	≥ 3	Yes	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper	breathes, can become ill, can't fly, contains proteins, dies, does not taste well, doesn't have wings, doesn't resemble a human being, eats, has a central nervous system, has brains, has eyes, has no hairs, has two eyes, is a living being, is able to reproduce, is an animal, is not a mammal, is useful, its vascular system is less developed than that of mammals, lays eggs, lives in the wild, lives in warm countries, uses the sun to keep body temperature stable
Reptile	≥ 4	No	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper, *salmon*, *cod*, *pike*, *plaice*, *goldfish*, *piranha*, *anchovy*, *ray*, *eel*, *trout*, *sole*, *sardine*, *stickleback*, *squid*, *shark*, *herring*, *carp*	breathes, can become ill, can't fly, dies, doesn't resemble a human being, eats, has a central nervous system, has eyes, has no hairs, has two eyes, is able to reproduce, is an animal, is not a mammal

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Clothing	≥ 1	Yes	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit	bought in a store, can also be used out of the kitchen, can be carried, can be recycled, can be seen on television, can be used inside and outside the house, can become dirty, can rip, can stand heat, can wear off, comes in different fabrics, comes in different materials, comes in different styles, comes in different weights, comes in very handy, comes into fashion/goes out of fashion, different types are available, differs from one culture to the other, doesn't have a specific odour, easily gets dirty, exists in different brands, exists in different colors, exists in different forms, exists in different kinds, exists in different sizes (groottes), exists in different sizes (maten), exists in different types, expresses your personality, has esthetic purposes, has no taste, implement/tool, is a collective noun, is a good invention, is available for each price range, is beautiful, is black, is bought in a specialist shop, is brown, is colourful, is dependent upon the occasion, is different for men and women, is flammable, is found in a cupboard, is functional, is light, is made by designers, is made by hand, is made in a factory, is made under bad circumstances, is meant to wear, is only used by humans, is pliable, is replaceable, is sewn, is shown on the catwalk, is sometimes given as a present for christmas, is used a lot, is used to cover yourself, needs good care, ready-to-wear clothes, sold in clothes shops
Clothing	≥ 2	Yes	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit	bought in a store, can also be used out of the kitchen, can be carried, can be seen on television, can be used inside and outside the house, can become dirty, can rip, can stand heat, can wear off, comes in different fabrics, comes in different materials, comes in different styles, comes into fashion/goes out of fashion, different types are available, doesn't have a specific odour, exists in different brands, exists in different colors, exists in different forms, exists in different kinds, exists in different sizes (groottes), exists in different sizes (maten), exists in different types, has esthetic purposes, is a good invention, is available for each price range, is dependent upon the occasion, is flammable, is found in a cupboard, is made be designers, is made in a factory, is made under bad circumstances, is meant to wear, is only used by humans, is replaceable, is sewn, is shown on the catwalk, ready-to-wear clothes
Clothing	≥ 3	Yes	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit	bought in a store, can also be used out of the kitchen, can become dirty, can stand heat, can wear off, different types are available, doesn't have a specific odour, exists in different brands, exists in different colors, exists in different kinds, exists in different types, is a good invention, is flammable, is found in a cupboard, is made be designers, is made in a factory, is only used by humans, is replaceable, is sewn, ready-to-wear clothes

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Clothing	≥ 4	No	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit, *plate*, *towel*, *mixer*, *toaster*	can become dirty, can wear off, different types are available, exists in different brands, exists in different colors, exists in different types, is a good invention, is only used by humans, is replaceable
Kitchen utensil	≥ 1	Yes	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok	bought in a store, can be dishwashed, can be recycled, can be seen on television, can become dirty, can stand heat, can wear off, comes in different materials, comes in different styles, comes in different weights, comes in very handy, different types are available, doesn't have a specific odour, easily gets dirty, exists in different brands, exists in different colors, exists in different forms, exists in different kinds, exists in different sizes (groottes), exists in different types, feels cold to the skin, has no taste, implement/tool, is a collective noun, is a good invention, is an aid, is available for each price range, is bought in a specialist shop, is easy to work with, is functional, is made be designers, is made in a factory, is only used by humans, is practical, is replaceable, is shared, is sometimes given as a present for christmas, is sometimes given to newly-weds as a present, is stainless, is the same for men and women, is used a lot, is used all over the world, is used to prepare food, is useful, makes work easier, makes work in the kitchen easier, used by cooks, used by everybody, used in the kitchen
Kitchen utensil	≥ 2	Yes	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok	bought in a store, can become dirty, can stand heat, can wear off, comes in different styles, comes in very handy, different types are available, doesn't have a specific odour, easily gets dirty, exists in different brands, exists in different forms, exists in different kinds, exists in different types, has no taste, implement/tool, is a good invention, is easy to work with, is functional, is made in a factory, is only used by humans, is practical, is replaceable, is shared, is useful, used in the kitchen
Kitchen utensil	≥ 3	Yes	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok	bought in a store, can become dirty, can stand heat, different types are available, exists in different brands, exists in different forms, is a good invention, is made in a factory, is only used by humans, is practical, is replaceable, is useful, used in the kitchen

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Kitchen utensil	≥ 4	No	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok, *harmonica*, *cymbals*, *file*, *rope*, *bagpipe*, *top*, *pullover*, *shorts*, *clamp*, *grinding disc*, *clarinet*, *oil can*, *bathing suit*, *nail*, *tongs*, *lawn mower*, *hammer*, *piano*, *adjustable spanner*, *blouse*, *sweater*, *flute*, *wire brush*, *skirt*, *dungarees*, *vacuum cleaner*, *chisel*, *cap*, *skateboard*, *double bass*, *bass guitar*, *level*, *tambourine*, *trumpet*, *saxophone*, *accordion*, *bra*, *synthesizer*, *paint brush*, *banjo*, *bassoon*, *wheelbarrow*, *jeans*, *shovel*, *drill*, *belt*, *harp*, *dress*, *beanie*, *trombone*, *mittens*, *hat*, *recorder*, *cello*, *kick scooter*, *t-shirt*, *pan flute*, *guitar*, *wrench*, *shirt*, *axe*, *motorbike (moto)*, *tracksuit*, *drum*, *bicycle*, *suit*, *tie*, *pickaxe*, *saw*, *coat*, *sled*, *plane*, *violin*, *screwdriver*, *pants*, *filling knife*, *drum set*, *anvil*, *crowbar (breekijzer)*, *scarf* , *bow*	bought in a store, different types are available, is a good invention, is only used by humans
Musical instrument	≥ 1	Yes	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin	bought in a store, can also be used out of the kitchen, can be dishwashed, can be played on, can be recycled, can be seen on television, can be used inside and outside the house, can be used to make music with, can become dirty, can move, can stand heat, can wear off, comes in different materials, comes in different styles, different types are available, doesn't have a specific odour, easily gets dirty, exists in different brands, exists in different colors, exists in different forms, exists in different kinds, exists in different sizes (grootte s), exists in different types, feels cold to the skin, for sale in a music shop, has no taste, implement/tool, is a collective noun, is a good invention, is available for each price range, is bought in a specialist shop, is firm, is found on a stage, is fun, is functional, is hard, is made be designers, is made by hand, is made in a factory, is meant to make music with, is only used by humans, is replaceable, is sometimes given as a present for christmas, is the same for men and women, is used all over the world, is used in orchestras, is used to communicate, is used to make something with, its vibrations produce sounds, makes a special noise, needs good care, one has to learn how to play, played by a musician, produces noise, produces sound, provides entertainment, provides you with freedom, sounds beautiful, you can earn money with it, you have to learn how to use it

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Musical instrument	≥ 2	Yes	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin	bought in a store, can also be used out of the kitchen, can be played on, can be seen on television, can be used inside and outside the house, can be used to make music with, can become dirty, can stand heat, can wear off, different types are available, doesn't have a specific odour, exists in different brands, exists in different colors, exists in different kinds, for sale in a music shop, is a good invention, is available for each price range, is bought in a specialist shop, is found on a stage, is made by hand, is meant to make music with, is only used by humans, is used in orchestras, its vibrations produce sounds, makes a special noise, needs good care, one has to learn how to play, played by a musician, produces noise, produces sound, provides entertainment, you can earn money with it, you have to learn how to use it
Musical instrument	≥ 3	Yes	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin	bought in a store, can also be used out of the kitchen, can be played on, can be used to make music with, can become dirty, different types are available, doesn't have a specific odour, exists in different brands, for sale in a music shop, is a good invention, is found on a stage, is meant to make music with, is only used by humans, one has to learn how to play, played by a musician, produces noise, produces sound
Musical instrument	≥ 4	Yes	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin	can be played on, exists in different brands, is a good invention, played by a musician, produces sound
Tool	≥ 1	Yes	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench	bought in a store, can also be used out of the kitchen, can be dishwashed, can be recycled, can be seen on television, can become dirty, can stand heat, can wear off, comes in different materials, comes in different styles, comes in different weights, comes in very handy, different types are available, doesn't have a specific odour, easily gets dirty, exists in different brands, exists in different colors, exists in different forms, exists in different kinds, exists in different sizes (grootes), exists in different sizes (maten), exists in different types, has no taste, implement/tool, is a collective noun, is a good invention, is a handtool, is a tool, is an aid, is available for each price range, is bought in a specialist shop, is easy to work with, is efficient, is functional, is made be designers, is made in a factory, is not for children, is only used by humans, is practical, is replaceable, is shared, is sometimes given as a present for christmas, is the same for men and women, is used all over the world, is used by a professional, is used to work with, is useful, makes work easier, used to work with

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Tool	≥ 2	Yes	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench	bought in a store, can also be used out of the kitchen, can be seen on television, can become dirty, can stand heat, can wear off, comes in different weights, comes in very handy, different types are available, doesn't have a specific odour, exists in different brands, exists in different colors, exists in different sizes (groottes), has no taste, implement/tool, is a good invention, is an aid, is bought in a specialist shop, is functional, is made in a factory, is only used by humans, is replaceable, is shared, is the same for men and women, is used to work with, is useful, makes work easier, used to work with
Tool	≥ 3	No	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench, *stove*, *sieve*, *colander*, *nutcracker*, *can opener*, *scissors*, *mixer*, *spatula*, *grater*, *whisk*	bought in a store, can become dirty, can stand heat, different types are available, implement/tool, is a good invention, is an aid, is made in a factory, is only used by humans, is replaceable, is the same for men and women, is useful, makes work easier
Tool	≥ 4	No	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench, *stove*, *tram*, *microwave oven*, *colander*, *bottle*, *kettle*, *fork*, *helicopter*, *shield*, *bowl*, *wok*, *plate*, *motorbike (brommer)*, *scooter*, *blouse*, *carriage*, *car*, *go-cart*, *bus*, *nutcracker*, *skateboard*, *hovercraft*, *fridge*, *train*, *truck (camion)*, *towel*, *bra*, *teaspoon*, *van*, *mug*, *spoon*, *percolator*, *jeep*, *sieve*, *can opener*, *tractor*, *airplane*, *scissors*, *scales*, *pan*, *mixer*, *bow*, *electric kettle*, *oven*, *apron*, *toaster*, *kick scooter*, *truck (vrachtwagen)*, *place mat*, *taxi*, *zeppelin*, *motorbike (moto)*, *cart*, *tie*, *bicycle*, *subway train*, *sled*, *pants*, *trailer*, *spatula*, *grater*, *pot*, *whisk*	is a good invention, is made in a factory, is only used by humans

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Vehicle	≥ 1	No	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin, *lawn mower*	bought in a store, can also be used out of the kitchen, can be dangerous, can be recycled, can be seen on television, can be used to get from one place to the other, can become dirty, can cause accidents, can move, can stand heat, can wear off, comes in different materials, comes in different styles, comes in different weights, consists of different parts, different types are available, doesn't have a specific odour, easily gets dirty, exists in different brands, exists in different colors, exists in different forms, exists in different kinds, exists in different sizes (groottes), exists in different sizes (maten), exists in different types, feels cold to the skin, has no taste, implement/tool, is a collective noun, is a faster form of transport than going by foot, is a good invention, is an aid, is available for each price range, is bought in a specialist shop, is firm, is functional, is grey, is hard, is made be designers, is made in a factory, is only used by humans, is replaceable, is shared, is subject to rules, is the same for men and women, its vibrations produce sounds, needs good care, you have to learn how to use it
Vehicle	≥ 2	No	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin, *drill*, *tank*	can also be used out of the kitchen, can be seen on television, can become dirty, can cause accidents, can move, can stand heat, can wear off, comes in different styles, different types are available, doesn't have a specific odour, exists in different brands, exists in different colors, exists in different forms, exists in different kinds, exists in different sizes (groottes), exists in different types, implement/tool, is made in a factory, is only used by humans, is replaceable
Vehicle	≥ 3	No	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin, *stove*, *microwave oven*, *colander*, *bottle*, *kettle*, *top*, *pullover*, *shorts*, *fork*, *shield*, *bowl*, *plate*, *bathing suit*, *lawn mower*, *blouse*, *sweater*, *glass*, *skirt*, *dungarees*, *nutcracker*, *sword*, *vacuum cleaner*, *cap*, *socks*, *fridge*, *towel*, *bra*, *teaspoon*, *mug*, *spoon*, *shoes*, *percolator*, *jeans*, *sieve*, *drill*, *belt*, *can opener*, *spear*, *dress*, *scissors*, *beanie*, *scales*, *pan*, *pistol*, *mixer*, *electric kettle*, *slingshot*, *hat*, *oven*, *apron*, *toaster*, *t-shirt*, *place mat*, *pyjamas*, *rifle*, *shirt*, *tracksuit*, *suit*, *panties*, *boots*, *tie*, *coat*, *pants*, *spatula*, *grater*, *pot*, *scarf*, *whisk*	can become dirty, can stand heat, different types are available, exists in different colors, exists in different forms, exists in different kinds, exists in different types, is made in a factory, is only used by humans

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Vehicle	≥ 4	No	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin, *stove*, *microwave oven*, *colander*, *kettle*, *top*, *pullover*, *shorts*, *fork*, *grinding disc*, *bowl*, *wok*, *plate*, *bathing suit*, *lawn mower*, *blouse*, *sweater*, *glass*, *skirt*, *dungarees*, *cap*, *socks*, *fridge*, *towel*, *bra*, *teaspoon*, *mug*, *spoon*, *paint brush*, *shoes*, *jeans*, *shovel*, *belt*, *plough*, *dress*, *beanie*, *pan*, *mittens*, *mixer*, *hat*, *toaster*, *t-shirt*, *pyjamas*, *shirt*, *tracksuit*, *suit*, *panties*, *boots*, *tie*, *coat*, *pants*, *spatula*, *pot*, *scarf* , *whisk*	can become dirty, different types are available, is only used by humans
Weapon	≥ 1	No	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip, *wheelbarrow*, *crowbar (koevoet)*, *shovel*, *drill*, *clamp*, *grinding disc*, *plough*, *knife*, *oil can*, *nail*, *tongs*, *lawn mower*, *chisel*, *wrench*, *saw*, *pickaxe*, *plane*, *screwdriver*, *filling knife*, *crowbar (breekijzer)*, *hammer*	can also be used out of the kitchen, can be dishwashed, can be recycled, can be seen on television, can become dirty, can stand heat, can wear off, comes in different materials, comes in different styles, comes in different weights, different types are available, doesn't have a specific odour, easily gets dirty, exists in different brands, exists in different colors, exists in different forms, exists in different kinds, exists in different sizes (grootte), exists in different sizes (maten), exists in different types, has no taste, implement/tool, is a collective noun, is an aid, is available for each price range, is dangerous for children, is efficient, is firm, is functional, is made in a factory, is not for children, is primarily used by men, is replaceable, is the same for men and women, is used all over the world, is useful
Weapon	≥ 2	No	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip, *file*, *jeep*, *submarine*, *crowbar (koevoet)*, *shovel*, *wheelbarrow*, *drill*, *clamp*, *grinding disc*, *helicopter*, *airplane*, *plough*, *knife*, *oil can*, *motorbike (brommer)*, *scooter*, *nail*, *tongs*, *lawn mower*, *hammer*, *adjustable spanner*, *boat*, *wire brush*, *car*, *truck (vrachtwagen)*, *taxi*, *zeppelin*, *vacuum cleaner*, *chisel*, *wrench*, *motorbike (moto)*, *cart*, *level*, *hovercraft*, *saw*, *pickaxe*, *subway train*, *plane*, *screwdriver*, *filling knife*, *rocket*, *anvil*, *crowbar (breekijzer)*, *paint brush*	can also be used out of the kitchen, can be seen on television, can become dirty, can stand heat, can wear off, different types are available, doesn't have a specific odour, exists in different brands, exists in different sizes (grootte), has no taste, implement/tool, is functional, is made in a factory, is replaceable, is the same for men and women

(*continued*)

Table A2. Continued

Category	Threshold	Is formal concept?	Extent	Intent
Weapon	≥ 3	No	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip, *harmonica, cymbals, stove, file, crowbar (koevoet), tram, microwave oven, colander, bottle, kettle, fork, clamp, grinding disc, helicopter, bowl, knife, wok, plate, oil can, motorbike (brommer), scooter, nail, tongs, lawn mower, hammer, adjustable spanner, wire brush, car, glass, go-cart, bus, nutcracker, vacuum cleaner, chisel, skateboard, level, hovercraft, tambourine, fridge, train, towel, rocket, teaspoon, van, mug, spoon, synthesizer, paint brush, percolator, jeep, sieve, submarine, wheelbarrow, shovel, drill, can opener, tractor, airplane, plough, scissors, triangle, scales, pan, mixer, electric kettle, oven, boat, recorder, toaster, kick scooter, truck (vrachtwagen), place mat, taxi, zeppelin, wrench, motorbike (moto), cart, drum, saw, pickaxe, sled, subway train, plane, screwdriver, filling knife, spatula, trailer, grater, drum set, crowbar (breekijzer), anvil, pot, whisk*	can become dirty, can stand heat, different types are available, is replaceable, is the same for men and women
Weapon	≥ 4	No	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip, *cymbals, stove, shorts, fork, clamp, grinding disc, helicopter, dungarees, nutcracker, chisel, fridge, train, accordion, truck (camion), spoon, jeep, sieve, submarine, plough, scales, pan, electric kettle, organ, place mat, wrench, panties, screwdriver, spatula, grater, anvil, scarf , harmonica, file, tram, microwave oven, bottle, top, pullover, plate, bathing suit, scooter, nail, tongs, lawn mower, piano, adjustable spanner, flute, go-cart, skateboard, double bass, level, tambourine, towel, bassoon, shovel, belt, dress, beanie, trombone, oven, toaster, cello, kick scooter, truck (vrachtwagen), pyjamas, taxi, guitar, shirt, motorbike (moto), cart, drum, boots, coat, sled, violin, trailer, drum set, whisk, colander, bagpipe, bowl, knife, wok, motorbike (brommer), car, glass, vacuum cleaner, socks, trumpet, harpsichord, synthesizer, shoes, wheelbarrow, jeans, drill, can opener, harp, tractor, airplane, scissors, boat, t-shirt, tracksuit, bicycle, pickaxe, plane, filling knife, crowbar (breekijzer), pot, kettle, clarinet, oil can, blouse, sweater, carriage, wire brush, skirt, bus, cap, bass guitar, hovercraft, saxophone, rocket, bra, teaspoon, van, mug, paint brush, banjo, percolator, mittens, mixer, hat, apron, recorder, zeppelin, pan flute, (hot air) balloon, suit, tie, saw, subway train, pants, hammer*	different types are available

Table A3. Categories and formal concepts of the animal and artifact domains with exemplar attributes.

Category	Threshold	Is formal concept?	Extent	Intent
Bird	≥ 1	Yes	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker	appears in fairy tales and stories, builds nests, can be aggressive, can be bred, can be caught, can be fed, can do without water for long periods, can turn his head very far, carries over diseases, communicates with its congeners, does not migrate in the winter, does not smell well, doesn't have 1000 paws, doesn't sting, drinks water, eats insects, eats small animals, eats worms, exists for ages, exists in different sizes and kinds, fragile, hairy (harig), has a bill, has a head, has a mouth, has a pointed mouth, has a sharp view, has a tail, has a tongue, has black eyes, has downy hair, has eyes, has feathers, has legs (poten), has round eyes, has sharp claws (nagels), has two ears, has two eyes, has two paws, has two wings, has wings, is a bird, is a carnivore, is an animal, is coloured, is difficult to catch, is eaten by other animals, is edible, is found in Belgium, is found on animals, is not dangerous, is not eaten, is not poisonous, is spectacular, jumps, lays eggs, lives in distant countries, lives in nature, lives in the open air, lives in the wild, lives on land, lives outdoors, makes a sound, neutral scent, occurs frequently, prey, small ears, smaller than a horse, some people are allergic to it, the meat is eaten, there are many kinds of it, you can boil it
Bird	≥ 2	Yes	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker	builds nests, can be aggressive, can be caught, can be fed, doesn't have 1000 paws, doesn't sting, drinks water, has a bill, has a head, has a mouth, has a tongue, has eyes, has feathers, has legs (poten), has round eyes, has two ears, has two eyes, has two paws, has two wings, has wings, is a bird, is an animal, is eaten by other animals, is not poisonous, lays eggs, lives in the open air, lives on land, lives outdoors, makes a sound, neutral scent, prey, small ears, smaller than a horse
Bird	≥ 3	Yes	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker	doesn't have 1000 paws, doesn't sting, drinks water, has a bill, has a head, has a tongue, has eyes, has legs (poten), has two eyes, has two paws, has wings, is a bird, is an animal, is not poisonous, lays eggs, makes a sound, smaller than a horse
Bird	≥ 4	Yes	blackbird, canary, chickadee, chicken, crow, cuckoo, dove, duck, eagle, falcon, heron, magpie, ostrich, owl, parakeet, parrot, peacock, pelican, penguin, pheasant, robin, rooster, seagull, sparrow, stork, swallow, swan, turkey, vulture, woodpecker	has a bill, has a head, has eyes, has legs (poten), has two eyes, has two paws, has wings, is a bird, is an animal, is not poisonous

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Fish	≥ 1	Yes	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale	appears in fairy tales and stories, bites, breaths under water, can be aggressive, can be bred, can be caught, can be fed, can swim, can't fly, communicates with its congeners, does not live in Belgium, does not migrate in the winter, doesn't have 1000 paws, doesn't live on land, doesn't make a sound, doesn't sting, drinks water, eats fish, eats plankton, eats small animals, eats worms, exists for ages, exists in different sizes and kinds, fragile, has a head, has a mouth, has a pungent smell, has a tail, has a tongue, has black eyes, has eyes, has fins, has gills, has no paws, has round eyes, has short paws, has teeth, has two ears, has two eyes, herds, is a carnivore, is an animal, is blue, is cold-blooded, is coloured, is difficult to catch, is eaten by other animals, is edible, is fast, is fished for, is found in Belgium, is found in warm places, is found on animals, is lithe, is not eaten, is not poisonous, is slimy, is slippery, is smooth, is sometimes eaten by man, is spectacular, likes humidity, lives by the sea, lives in Africa, lives in America, lives in Europe, lives in a damp climate, lives in cold areas, lives in distant countries, lives in nature, lives in the ocean, lives in the open air, lives in the sea, lives in the wild, lives in warm countries, lives in water, lives nearby the water, lives outdoors, makes not much noise, occurs frequently, prey, produces blub-sound, some people are allergic to it, stinks, the meat is eaten, there are many kinds of it, wild animal, you can boil it
Fish	≥ 2	Yes	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale	breaths under water, can be caught, can be fed, can swim, can't fly, does not migrate in the winter, doesn't have 1000 paws, doesn't live on land, doesn't sting, drinks water, eats small animals, has a head, has a mouth, has a tongue, has black eyes, has eyes, has fins, has gills, has no paws, has teeth, has two eyes, is an animal, is eaten by other animals, is edible, is not poisonous, is slimy, is slippery, is smooth, likes humidity, lives in a damp climate, lives in nature, lives in the open air, lives in the sea, lives in the wild, lives in warm countries, lives in water, lives outdoors, makes not much noise, prey, the meat is eaten

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Fish	≥ 3	Yes	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale	can swim, can't fly, doesn't have 1000 paws, doesn't live on land, doesn't sting, drinks water, has a head, has eyes, has gills, has no paws, has two eyes, is an animal, is not poisonous, is slippery, is smooth, lives in nature, lives in the wild, lives in water
Fish	≥ 4	Yes	anchovy, carp, cod, dolphin, eel, flatfish, goldfish, herring, orca, pike, piranha, plaice, ray, salmon, sardine, shark, sole, sperm whale, squid, stickleback, swordfish, trout, whale	can swim, can't fly, has a head, has eyes, has two eyes, is an animal, is slippery, is smooth, lives in nature, lives in water
Insect	≥ 1	Yes	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm	Indians or African, also lives in the city, appears in fairy tales and stories, builds nests, can be aggressive, can be bred, can be caught, can be found in the ardennes, can become pest, can do without water for long periods, communicates with its congeners, does not migrate in the winter, does not taste well, does not have 1000 paws, drinks water, exists for ages, exists in different sizes and kinds, fragile, has a head, has a mouth, has a small head, has a tongue, has black eyes, has eyes, has feelers, has little eyes, has round eyes, has short paws, has teeth, has two eyes, herds, is an animal, is an articulate animal, is an insect, is athrodopal, is coloured, is eaten by birds, is eaten by other animals, is edible, is exterminated, is food for larger animals, is found in Belgium, is found in the garden, is found in warm places, is found on animals, is inconspicious, is killed by man, is light, is not a pet, is not eaten, is not expensive, is not poisonous, is not very big, is not very popular among man, is small (klein), is spectacular, lays eggs, lives by the sea, lives in Africa, lives in America, lives in Europe, lives in Australia, lives in distant countries, lives in fields, lives in nature, lives in the jungle, lives in the open air, lives in the wild, lives in the woods, lives in tropical areas, lives in warm countries, lives on land, lives outdoors, makes not much noise, neutral scent, occurs frequently, often sits in parks, prey, reproduces fast, small ears, smaller than a horse, some people are allergic to it, the meat is eaten, there are many kinds of it, you can boil it

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Insect	≥ 2	Yes	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm	can be caught, does not migrate in the winter, does not taste well, does not have 1000 paws, drinks water, has a head, has a mouth, has eyes, has feelers, has short paws, has two eyes, is an animal, is an insect, is eaten by other animals, is found in Belgium, is found in the garden, is light, is not a pet, is not eaten, is not poisonous, is not very big, is small (klein), lays eggs, lives in Africa, lives in America, lives in Europe, lives in nature, lives in the jungle, lives in the open air, lives in the wild, lives in warm countries, lives on land, lives outdoors, neutral scent, prey, smaller than a horse
Insect	≥ 3	Yes	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm	does not taste well, does not have 1000 paws, drinks water, has a head, is an animal, is an insect, is found in Belgium, is light, is not a pet, is not very big, is small (klein), lays eggs, lives in Europe, lives in nature, lives in the open air, lives in the wild, lives on land, lives outdoors, smaller than a horse
Insect	≥ 4	Yes	ant, bee, beetle, bumblebee, butterfly, caterpillar, centipede, cockchafer, cockroach, cricket, dragonfly, earwig, flee, fly, fruit fly, grasshopper, horsefly, ladybug, leech, louse, mosquito, moth, spider, wasp, wood louse, worm	is an animal, is found in Belgium, is light, is not very big, is small (klein), lives in Europe, lives in the open air, lives on land
Mammal	≥ 1	Yes	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra	can be aggressive, can be bred, can be caught, can be fed, can do without water for long periods, communicates with its congeners, does not lay eggs, does not migrate in the winter, does not smell well, does not have 1000 paws, does not sting, drinks water, exists for ages, exists in different sizes and kinds, fragile, gives milk, good sense of smell, has a fur (vacht), has a head, has a mouth, has a tail, has a tongue, has eyes, has four paws, has legs (poten), has many teeth, has round eyes, has teeth, has two ears, has two eyes, is a carnivore, is an animal, is coloured, is eaten by other animals, is edible, is found in Belgium, is found on animals, is not eaten, is not poisonous, is spectacular, lives in Africa, lives in distant countries, lives in the open air, lives on land, lives on the ground, lives outdoors, makes a sound, mammal, prey, some people are allergic to it, the meat is eaten, there are many kinds of it, thick skin, you can boil it

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Mammal	≥ 2	Yes	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra	can be fed, does not lay eggs, does not migrate in the winter, does not have 1000 paws, does not sting, drinks water, gives milk, has a head, has a mouth, has a tail, has a tongue, has eyes, has legs (poten), has round eyes, has teeth, has two ears, has two eyes, is an animal, is eaten by other animals, is not poisonous, lives in the open air, lives on land, lives outdoors, makes a sound, mammal
Mammal	≥ 3	Yes	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra	does not lay eggs, does not have 1000 paws, does not sting, drinks water, has a head, has a mouth, has a tongue, has eyes, has legs (poten), has teeth, has two ears, has two eyes, is an animal, is not poisonous, lives on land, mammal
Mammal	≥ 4	Yes	bat, beaver, bison, cat, cow, deer, dog, donkey, dromedary, elephant, fox, giraffe, hamster, hedgehog, hippopotamus, horse, kangaroo, lion, llama, monkey, mouse, pig, polar bear, rabbit, rhinoceros, sheep, squirrel, tiger, wolf, zebra	has a tongue, has eyes, has two ears, has two eyes, is an animal, is not poisonous, mammal
Reptile	≥ 1	Yes	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper	Indians or African, bites, builds nests, can be aggressive, can be bred, can be caught, can be fed, can do without water for long periods, can live for a long time without food, can swim, cannot fly, communicates with its congeners, dark colour, does not migrate in the winter, does not taste well, does not have 1000 paws, does not herd, does not sting, drinks water, eats insects, eats plants, eats small animals, eats worms, exists for ages, exists in different sizes and kinds, fragile, has a head, has a mouth, has a tail, has a tongue, has black eyes, has eyes, has round eyes, has scales, has short paws, has teeth, has two eyes, is a carnivore, is a reptile, is an animal, is cold-blooded, is coloured, is eaten by other animals, is found in warm places, is found mainly in southern countries, is found on animals, is green-brown, is not a pet, is not eaten, is spectacular, is ugly, lays eggs, lives in Africa, lives in America, lives in distant countries, lives in grass, lives in nature, lives in the jungle, lives in the open air, lives in the wild, lives in the zoo, lives in tropical areas, lives in warm countries, lives nearby the water, lives on land, lives on the ground, lives outdoors, makes not much noise, neutral scent, occurs frequently, prey, sheds its skin, shy, smaller than a horse, there are many kinds of it, wild animal, you can boil it

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Reptile	≥ 2	Yes	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper	can be aggressive, can be caught, can be fed, cannot fly, does not migrate in the winter, does not taste well, does not have 1000 paws, does not sting, drinks water, eats insects, eats small animals, has a head, has a mouth, has a tongue, has eyes, has round eyes, has two eyes, is a reptile, is an animal, is cold-blooded, is found mainly in southern countries, is not a pet, lays eggs, lives in Africa, lives in nature, lives in the open air, lives in the wild, lives in warm countries, lives outdoors, neutral scent, prey, smaller than a horse
Reptile	≥ 3	Yes	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper	cannot fly, does not taste well, does not have 1000 paws, does not sting, drinks water, has a head, has a tongue, has eyes, has two eyes, is an animal, lays eggs, lives in nature, lives in the open air, lives in the wild, lives in warm countries, lives outdoors
Reptile	≥ 4	No	alligator, blindworm, boa, caiman, chameleon, cobra, crocodile, dinosaur, frog, gecko, iguana, lizard, monitor lizard, python, salamander, snake, toad, tortoise, turtle, viper, *hippopotamus*, *beaver*, *deer*, *lion*, *mouse*, *kangaroo*, *polar bear*, *wolf* , *elephant*, *squirrel*, *tiger*, *rhinoceros*, *giraffe*, *fox*, *hedgehog*, *zebra*	cannot fly, has a head, has a tongue, has eyes, has two eyes, is an animal, lives in nature, lives in the open air

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Clothing	≥ 1	Yes	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit	both large and small, bought in a store, can be carried, can be coloured, can be decorated, can be held, can be printed with a design, can be seen on television, can be used several times, can become dirty, can break (stuk gaan), can sink, comes in different brands and models, comes in different patterns, comes in different prints, comes in very handy, constitutes a whole, costs money, does not produce sound, does not weigh much, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different forms, exists in different kinds, exists in different materials, exists in different measurements, exists in different sizes (groottes), exists in different sizes (maten), exists in different types, has a dark color, has a plain color, has different patterns (motiefjes), has no engine, has no roof, has no smell, has no taste, has the name of a company printed on it, implement/tool, is a garment, is an object, is available for each price range, is beautiful, is black, is brown, is clothing, is colourful, is for 1 person, is functional, is handy to transport, is invented by man, is lifeless, is light, is made in a factory, is not edible, is not waterproof, is odourless, is pliable, is smaller than a lorry, is used to cover yourself, lies in the cupboard, made of fibres, made of synthetic material, not everyone has it, occurs in comic books, occurs in films, sold in clothes shops, the size is indicated by numbers, used in different cultures, was used in the past, water permeable, worn by people, you do not have to peddle, you have to pay for it

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Clothing	≥ 2	Yes	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit	bought in a store, can be carried, can be coloured, can be held, can be seen on television, can be used several times, can become dirty, can break (stuk gaan), can sink, comes in different brands and models, comes in different patterns, comes in different prints, costs money, does not produce sound, does not weigh much, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different forms, exists in different kinds, exists in different sizes (grootes), exists in different sizes (maten), exists in different types, has no engine, has no roof, has no smell, is a garment, is an object, is available for each price range, is clothing, is for 1 person, is invented by man, is lifeless, is made in a factory, is not edible, is odourless, is smaller than a lorry, not everyone has it, the size is indicated by numbers, worn by people, you have to pay for it
Clothing	≥ 3	Yes	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit	bought in a store, can be coloured, can be held, can become dirty, can break (stuk gaan), comes in different brands and models, costs money, does not produce sound, does not weigh much, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different kinds, exists in different types, is a garment, is an object, is clothing, is for 1 person, is invented by man, is lifeless, is made in a factory, is not edible, worn by people
Clothing	≥ 4	Yes	bathing suit, beanie, belt, blouse, boots, bra, cap, coat, dress, dungarees, hat, jeans, mittens, panties, pants, pullover, pyjamas, scarf, shirt, shoes, shorts, skirt, socks, suit, sweater, t-shirt, tie, top, tracksuit	can become dirty, comes in different brands and models, does not produce sound, exists in different brands, exists in different colors, exists in different types, is a garment

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Kitchen utensil	≥ 1	Yes	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok	both large and small, bought in a store, can be dishwashed, can be made of different materials, can be made of different tissues, can be seen on television, can be used by anyone, can be used several times, can become dirty, can break (stuk gaan), can sink, comes in different brands and models, comes in very handy, constitutes a whole, costs money, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different forms, exists in different kinds, exists in different sizes (groottes), exists in different types, has a plain color, has no engine, has no roof, has no smell, has no taste, implement/tool, is an object, is available for each price range, is easy, is easy to work with, is for at least 1 person, is functional, is invented by man, is lifeless, is made in a factory, is not edible, is not sexy, is odourless, is smaller than a lorry, is stainless, is useful, needs to be cleaned sometimes, not everyone has it, occurs in comic books, occurs in films, used by cooks, used by everybody, used by men, used by mothers, used in a restaurant, used in different cultures, used in the house, used in the kitchen, used with your hands, was used in the past, you don't have to peddle, you have to pay for it
Kitchen utensil	≥ 2	Yes	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok	bought in a store, can be used several times, can become dirty, can break (stuk gaan), can sink, comes in different brands and models, comes in very handy, costs money, exists in different brands, exists in different forms, exists in different kinds, exists in different types, has no roof, has no smell, has no taste, implement/tool, is an object, is easy to work with, is functional, is invented by man, is lifeless, is made in a factory, is not edible, is not sexy, is odourless, is smaller than a lorry, is useful, not everyone has it, used in different cultures, used in the house, used in the kitchen
Kitchen utensil	≥ 3	Yes	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok	bought in a store, can be used several times, can become dirty, comes in different brands and models, costs money, exists in different brands, exists in different forms, is an object, is invented by man, is made in a factory, is not edible, is not sexy, is useful, used in the house, used in the kitchen

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Kitchen utensil	≥ 4	No	apron, bottle, bowl, can opener, colander, electric kettle, fork, fridge, glass, grater, kettle, knife, microwave oven, mixer, mug, nutcracker, oven, pan, percolator, place mat, plate, pot, scales, scissors, sieve, spatula, spoon, stove, teaspoon, toaster, towel, whisk, wok, *harmonica*, *cymbals*, *file*, *knuckle dusters*, *rope*, *crowbar (koevoet)*, *bagpipe*, *clamp*, *grinding disc*, *clarinet*, *oil can*, *bathing suit*, *nail*, *tongs*, *lawn mower*, *hammer*, *piano*, *adjustable spanner*, *flute*, *wire brush*, *vacuum cleaner*, *chisel*, *double bass*, *bass guitar*, *level*, *tambourine*, *trumpet*, *saxophone*, *accordion*, *dagger*, *harpsichord*, *synthesizer*, *paint brush*, *banjo*, *bassoon*, *wheelbarrow*, *shovel*, *drill*, *harp*, *triangle*, *trombone*, *club*, *recorder*, *cello*, *pan flute*, *guitar*, *wrench*, *axe*, *drum*, *tie*, *saw*, *pickaxe*, *violin*, *plane*, *screwdriver*, *filling knife*, *drum set*, *anvil*, *crowbar (breekijzer)*, *bow*, *whip*	bought in a store, is an object
Musical instrument	≥ 1	Yes	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin	accompanies (other) music, both large and small, bought in a store, can be decorated, can be dishwashed, can be made of different materials, can be played on, can be played upon, can be seen on television, can be used by anyone, can be used several times, can become dirty, can break (stuk gaan), can sink, comes in different brands and models, constitutes a whole, costs money, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different forms, exists in different kinds, exists in different sizes (grootes), exists in different types, for sale in a music shop, has a plain color, has no engine, has no roof, has no smell, has no taste, has the name of a company printed on it, implement/tool, is a musical instrument, is a piece of equipment, is an object, is available for each price range, is firm, is for 1 person, is for all ages, is for few persons, is for fun, is fun, is functional, is hard, is hard to iron, is invented by man, is lifeless, is made in a factory, is not edible, is not sexy, is odourless, is played during concerts, is relaxing, is smaller than a lorry, its vibrations produce sounds, makes a nice sound/noise, makes a special noise, needs to be cleaned sometimes, not everyone has it, occurs in comic books, occurs in films, one has to learn how to play, played by a musician, played by a single person, played with the hands, played with two hands, produces music, produces noise, produces sound, smells neutral, used in different cultures, used solo, used with your hands, using it is fun, vibrates, was used in the past, you can learn it in a school of music, you can play with it, you don't have to peddle, you have to pay for it, you have to read music-scores

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Musical instrument	≥ 2	Yes	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin	bought in a store, can be played on, can be played upon, can be seen on television, can be used several times, can become dirty, can break (stuk gaan), comes in different brands and models, costs money, exists in different brands, exists in different colors, exists in different kinds, for sale in a music shop, has no engine, has no smell, is a musical instrument, is an object, is available for each price range, is for 1 person, is for all ages, is for fun, is invented by man, is lifeless, is not edible, is not sexy, is played during concerts, is smaller than a lorry, its vibrations produce sounds, makes a nice sound/noise, makes a special noise, not everyone has it, one has to learn how to play, played by a musician, played by a single person, produces music, produces noise, produces sound, used with your hands, vibrates, you can play with it, you have to pay for it
Musical instrument	≥ 3	Yes	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin	bought in a store, can be played on, can be played upon, can become dirty, can break (stuk gaan), comes in different brands and models, costs money, exists in different brands, for sale in a music shop, has no engine, is a musical instrument, is an object, is invented by man, is lifeless, is not edible, is played during concerts, makes a nice sound/noise, not everyone has it, one has to learn how to play, played by a musician, played by a single person, produces music, produces noise, produces sound
Musical instrument	≥ 4	Yes	accordion, bagpipe, banjo, bass guitar, bassoon, cello, clarinet, cymbals, double bass, drum, drum set, flute, guitar, harmonica, harp, harpsichord, organ, pan flute, piano, recorder, saxophone, synthesizer, tambourine, triangle, trombone, trumpet, violin	can be played on, costs money, exists in different brands, is an object, played by a musician, produces sound
Tool	≥ 1	Yes	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench	both large and small, bought in a store, can be dishwashed, can be seen on television, can be used several times, can become dirty, can break (stuk gaan), can sink, comes in different brands and models, comes in very handy, constitutes a whole, costs money, especially used by boys, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different forms, exists in different kinds, exists in different measurements, exists in different sizes (groottes), exists in different sizes (maten), exists in different types, has a plain color, has no roof, has no smell, has no taste, implement/tool, is a handtool, is a piece of equipment, is a tool, is an object, is available for each price range, is durable, is easy to work with, is efficient, is for 1 person, is for at least 1 person, is for few persons, is for sturdy guys, is found in a workplace, is functional, is hard to iron, is invented by man, is lifeless, is made in a factory, is not edible, is not for children, is not sexy, is odourless, is smaller than a lorry, is useful, needs to be cleaned sometimes, not everyone has it, occurs in comic books, occurs in films, smells neutral, used by men, used by the handyman, used by work forces, used in different cultures, used to work with, used with your hands, was used in the past, you don't have to peddle

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Tool	≥ 2	Yes	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench	bought in a store, can be seen on television, can be used several times, can become dirty, can break (stuk gaan), comes in different brands and models, comes in very handy, constitutes a whole, costs money, exists in different brands, exists in different colors, exists in different sizes (groottes), has no roof, has no smell, has no taste, implement/tool, is a piece of equipment, is an object, is functional, is invented by man, is lifeless, is made in a factory, is not edible, is not sexy, is odourless, is smaller than a lorry, is useful, not everyone has it, used by work forces, used to work with
Tool	≥ 3	No	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench, *stove*, *sieve*, *kettle*, *colander*, *can opener*, *fork*, *scissors*, *wok*, *pan*, *mixer*, *glass*, *nutcracker*, *fridge*, *spatula*, *grater*, *teaspoon*, *pot*, *whisk*, *spoon*	bought in a store, can be used several times, can become dirty, costs money, implement/tool, is an object, is made in a factory, is not edible, is not sexy, is useful
Tool	≥ 4	No	adjustable spanner, anvil, axe, chisel, clamp, crowbar (breekijzer), crowbar (koevoet), drill, file, filling knife, grinding disc, hammer, knife, lawn mower, level, nail, oil can, paint brush, pickaxe, plane, plough, rope, saw, screwdriver, shovel, tongs, vacuum cleaner, wheelbarrow, wire brush, wrench, *stove*, *microwave oven*, *kettle*, *bottle*, *colander*, *canon*, *fork*, *shield*, *bowl*, *plate*, *wok*, *nutcracker*, *fridge*, *towel*, *teaspoon*, *mug*, *spoon*, *percolator*, *sieve*, *can opener*, *grenade*, *scissors*, *bazooka*, *scales*, *pan*, *pistol*, *mixer*, *bow*, *electric kettle*, *oven*, *apron*, *toaster*, *double-barreled shotgun*, *place mat*, *rifle*, *machine gun*, *tie*, *spatula*, *grater*, *knuckle dusters*, *pot*, *whisk*	is an object, is made in a factory

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Vehicle	≥ 1	Yes	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin	both large and small, bought in a store, can be coloured, can be made of different materials, can be made of different tissues, can be seen on television, can be used several times, can become dirty, can break (breken), can break (stuk gaan), can crash, can sink, comes in different brands and models, consists of different parts, constitutes a whole, costs money, does not melt in high temperature, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different forms, exists in different kinds, exists in different lengths, exists in different materials, exists in different measurements, exists in different sizes (groottes), exists in different sizes (maten), exists in different types, has a plain color, has no smell, has no taste, has the name of a company printed on it, implement/tool, is a means of transport, is a means of transportation, is a vehicle, is an object, is available for each price range, is durable, is firm, is for at least 1 person, is functional, is grey, is handy to transport, is hard, is hard to iron, is invented by man, is lifeless, is made in a factory, is manned, is not edible, is not sexy, is odourless, is strong, is to be used outside, it has screws in it, its vibrations produce sounds, needs to be cleaned sometimes, not everyone has it, occurs in comic books, occurs in films, smells neutral, transports persons or goods, used in different cultures, vibrates, was used in the past, you have to pay for it
Vehicle	≥ 2	No	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin, *pistol*	can be coloured, can be made of different materials, can be seen on television, can be used several times, can become dirty, can break (stuk gaan), comes in different brands and models, costs money, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different forms, exists in different kinds, exists in different materials, exists in different sizes (groottes), exists in different types, implement/tool, is an object, is invented by man, is lifeless, is made in a factory, is not edible, is not sexy, is to be used outside, it has screws in it, needs to be cleaned sometimes, not everyone has it

(continued)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Vehicle	$\geq$3	No	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin, *stove*, *microwave oven*, *colander*, *top*, *shield*, *bathing suit*, *lawn mower*, *blouse*, *sweater*, *skirt*, *dungarees*, *nutcracker*, *sword*, *cap*, *fridge*, *bra*, *percolator*, *sieve*, *drill*, *belt*, *can opener*, *spear*, *dress*, *scales*, *pistol*, *mixer*, *electric kettle*, *slingshot*, *hat*, *toaster*, *rifle*, *shirt*, *suit*, *tracksuit*, *boots*, *spatula*, *whisk*	can be made of different materials, can be used several times, can become dirty, can break (stuk gaan), costs money, exists in different colors, exists in different forms, exists in different kinds, exists in different types, is an object, is invented by man, is made in a factory, is not edible, not everyone has it
Vehicle	$\geq$4	No	(hot air) balloon, airplane, bicycle, boat, bus, car, carriage, cart, go-cart, helicopter, hovercraft, jeep, kick scooter, motorbike (brommer), motorbike (moto), rocket, scooter, skateboard, sled, submarine, subway train, taxi, tractor, trailer, train, tram, truck (camion), truck (vrachtwagen), van, zeppelin, *stove*, *grinding disc*, *plate*, *fridge*, *lawn mower*, *toaster*, *glass*, *paint brush*	can become dirty, can break (stuk gaan)
Weapon	$\geq$1	No	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip, *wheelbarrow*, *shovel*, *chisel*, *plough*, *saw*, *plane*, *screwdriver*, *tongs*, *hammer*	both large and small, can be dishwashed, can be made of different materials, can be seen on television, can be used several times, can become dirty, can break (breken), can break (stuk gaan), can sink, comes in different brands and models, constitutes a whole, costs money, exists in different brands, exists in different colors, exists in different colors and different forms, exists in different forms, exists in different kinds, exists in different measurements, exists in different sizes (grootes), exists in different sizes (maten), exists in different types, has a dark color, has a plain color, has no engine, has no roof, has no smell, has no taste, implement/tool, is a piece of equipment, is an object, is available for each price range, is dangerous for children, is durable, is efficient, is firm, is for at least 1 person, is for few persons, is for sturdy guys, is functional, is hard to iron, is invented by man, is lifeless, is made in a factory, is not edible, is not for children, is not sexy, is odourless, is smaller than a lorry, is strong, is useful, needs to be cleaned sometimes, not everyone has it, occurs in comic books, occurs in films, smells neutral, used by men, used by strong men, used in different cultures, was used in the past, you don't have to peddle

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Weapon	≥ 2	No	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip, *file, wheelbarrow, crowbar (koevoet), shovel, drill, clamp, grinding disc, knife, oil can, nail, tongs, adjustable spanner, wire brush, chisel, wrench, level, saw, pickaxe, plane, screwdriver, crowbar (breekijzer), anvil, hammer, paint brush*	can be seen on television, can be used several times, can become dirty, can break (stuk gaan), constitutes a whole, costs money, exists in different brands, exists in different sizes (groottes), has a plain color, has no smell, has no taste, implement/tool, is an object, is functional, is lifeless, is made in a factory, is not edible, is not sexy, is odourless, is smaller than a lorry, not everyone has it, smells neutral, used by men
Weapon	≥ 3	No	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip, *harmonica, cymbals, stove, file, tram, microwave oven, colander, bagpipe, bottle, top, pullover, shorts, clamp, tie, grinding disc, helicopter, clarinet, bowl, knife, plate, oil can, bathing suit, motorbike (brommer), scooter, nail, tongs, lawn mower, hammer, piano, blouse, sweater, carriage, flute, car, glass, skirt, go-cart, dungarees, bus, nutcracker, vacuum cleaner, chisel, cap, skateboard, socks, double bass, bass guitar, level, hovercraft, trumpet, tambourine, saxophone, fridge, train, accordion, truck (camion), rocket, bra, van, harpsichord, mug, synthesizer, paint brush, banjo, shoes, bassoon, percolator, jeep, jeans, sieve, submarine, wheelbarrow, shovel, drill, belt, can opener, harp, tractor, airplane, plough, dress, scissors, triangle, beanie, scales, pan, trombone, mittens, electric kettle, mixer, hat, oven, boat, recorder, toaster, cello, organ, kick scooter, truck (vrachtwagen), t-shirt, pyjamas, taxi, zeppelin, pan flute, guitar, (hot air) balloon, shirt, motorbike (moto), tracksuit, panties, bicycle, cart, drum, pickaxe, boots, suit, saw, plane, coat, sled, subway train, violin, screwdriver, pants, filling knife, spatula, trailer, drum set, pot, scarf , whisk*	can become dirty, can break (stuk gaan), is an object, is not edible

(*continued*)

Table A3. Continued.

Category	Threshold	Is formal concept?	Extent	Intent
Weapon	≥ 4	No	axe, bazooka, bow, canon, club, dagger, double-barreled shotgun, grenade, knuckle dusters, machine gun, pistol, rifle, rope, shield, slingshot, spear, stick, sword, tank, whip, *cymbals, stove, shorts, fork, clamp, grinding disc, helicopter, dungarees, nutcracker, chisel, fridge, train, accordion, truck (camion), spoon, jeep, sieve, submarine, plough, scales, pan, electric kettle, organ, place mat, wrench, panties, screwdriver, spatula, grater, anvil, scarf , harmonica, file, crowbar (koevoet), tram, microwave oven, bottle, top, pullover, plate, bathing suit, scooter, nail, tongs, lawn mower, piano, adjustable spanner, flute, go-cart, skateboard, double bass, level, tambourine, towel, bassoon, shovel, belt, dress, beanie, trombone, oven, toaster, cello, kick scooter, truck (vrachtwagen), pyjamas, taxi, guitar, shirt, motorbike (moto), cart, drum, boots, coat, sled, violin, trailer, drum set, whisk, colander, bagpipe, bowl, knife, wok, motorbike (brommer), car, glass, vacuum cleaner, socks, trumpet, harpsichord, synthesizer, shoes, wheelbarrow, jeans, drill, can opener, harp, tractor, airplane, scissors, triangle, boat, t-shirt, tracksuit, bicycle, pickaxe, plane, filling knife, crowbar (breekijzer), pot, kettle, clarinet, oil can, blouse, sweater, carriage, wire brush, skirt, bus, cap, bass guitar, hovercraft, saxophone, rocket, bra, teaspoon, van, mug, paint brush, banjo, percolator, mittens, mixer, hat, apron, recorder, zeppelin, pan flute, (hot air) balloon, suit, tie, saw, subway train, pants, hammer*	