Category structure of language types common to conceptual modeling languages

Abcd efg hij Klmnop[1,2,3] and Qrstuvwx Yzabcd[1,2,3]

[1] Institute 1, City, Country
{abcd.efghijklmnop,qrst.uvwxyz}@abcde.fg
[2] Institute 2, City, Country
[3] Institute 3, City, Country*

Abstract. We investigate the category structure of categories common to conceptual modeling languages (i.e., the types used by languages such as actor, process, goal, or restriction) to study whether they more closely approximate a discrete or graded category. We do this for three distinct groups: students, beginning modelers and experienced modelers. We find that overall most categories exhibit more of a graded structure, with experienced modelers displaying this even more strongly than the other groups. We discuss the consequences of these results for (conceptual) modeling in general, and in particular argue that when a model contains graded categories, it should follow that the (conceptual) validity of instantiations of it should be judged in a graded fashion as well.

Keywords: categorization, conceptual modeling, model semantics

1 Introduction

We categorize the world around us in different ways depending on the subject matter. Some things we categorize more discretely, like natural things (e.g., fruits and plants), some things we categorize in a more graded way, such as artifactual things (e.g., tools, vehicles). These different categorization tendencies have been shown many times in research, starting around the time of Rosch et al. [22, 23]. Also, they have been investigated by many others explicitly elaborating on the category structure for a number of natural and artifactual categories (cf. [8, 4, 9, 10]). On the other hand, some work investigating this has had difficulties in finding significant differences in categorization tendencies between artifactual and natural categories (cf. [17]). There are also arguments that the natural/artifactual distinction is not granular enough, requiring us to also distinguish emotion categories [3]. Regardless of the debate whether particular kinds of categories are usually categorized in a particular way, it is clear that *we do not categorize everything in the same way.*

* Mandatory disclaimer about institute 3 which normally takes up like the three first lines so this filler should too so all the tables don't shift and break. Boy, you sure need to type a lot to replace three lines with anonymous boulderdash.(`www.website.url`)

The categorization we speak of here deals with membership judgments. That is, whether a certain thing is judged to be a member of a given category. For example, most people would have no problem saying that an apple is a member of the category FRUIT[4], and they will likely reject the notion of a newspaper being so. However, when borderline cases are introduced interesting effects occur [13, 14]. Given, for instance, cases that do not have clear or crisp boundaries, like tomatoes or rhubarb, people have more difficulty deciding with certainty whether they are FRUITS or not. In such cases people often tend to give *graded judgments* – things being members of a category to a certain degree. This prevalence of (strongly) graded membership judgments is then often correlated with the structure of the category being graded. Given that many of our modeling efforts (be they the creation of domain models, ontologies to formalize knowledge or support reasoning with, databases to implement schemata, etc.) require us to be as exact as possible about what we aim to model, it is clear that being aware of such differences in membership judgments is an important aspect of properly representing a given domain and the things in it.

The importance of being aware of these different judgments starts during the modeling phase, particularly in settings where there is collaborative modeling and integration efforts (e.g., enterprise modeling). The uncertainty of membership judgments (i.e., what is a valid instantiation for this type, is this instantiation as valid as others) creeps into models, and is often lost, unless explicitly elicited and written down. The effect this has on the validity of a model can occur on two levels, the level of the categories from the domain (i.e., the concepts from the universe of discourse) and the level of the categories from the language (i.e., the types used by a modeling language). Domain categories – the concepts from the universe of discourse – often receive great attention in discussions between modelers and stakeholders as well as in discussion between modelers themselves. This ensures (to some degree) that modelers know what things the stakeholders want to see in a model, and that they understand those things in the same way [16]. However, categories from the language receive such detailed attention far less often, e.g., by asking "*What exactly is this type 'actor' from the language we are using? Does it allow us to model the acting elements from the universe of discourse we know about?*". Instead, we often end up using the semantics of our own natural language [25] – together with all the category structures and nuances that come with it. Because of this, the language that ends up actually being used often differs from the (formal definition of the) modeling language that is used on paper [15]. For example, a modeling language might formally define an actor as a rather specific thing (e.g., requiring it to be a singular abstract entity, and whatever other features might apply), which makes it fairly easy to determine whether something is a valid instantiation of that type – a human being here definitely not being one. On the other hand, one of the modelers (or any reader of the model) might not use (or indeed, be aware of) those semantics, and instead see the type as having a different range of conceptually valid instantiations. This is problematic because it means that important semantics of the

[4] To distinguish categories from words we print them in SMALL CAPS.

model might be lost when it is interpreted by other people not involved in the original modeling process (e.g., during model integration), or stakeholders who were not aware of some of the not explicated particularities. This is exacerbated by the fact that we do not have an insight into the structure of these categories *as used* by people, because not only do we not know what is considered valid, we do not know whether some things are considered more valid than others.

Thus, in this paper we aim to clarify whether the categories common to many modeling languages and methods (i.e., those types used by a language to instantiate domain concepts by) are categorized in a discrete or graded fashion. The implications of this for model creation and usage (particularly for models used to capture and document a certain domain) are important to be aware of. If a category from a language is typically judged in a discrete fashion, the semantics of models are likely easier to communicate, formalize, and keep coherent. However, if such a category is typically judged in a graded fashion, communicating it to others becomes more involved, requiring more explicit discussion, and the formalizations and tools we use need to explicitly support this structure (e.g., by using ontologies with support for features as typicality and centrality).

To the best of our knowledge there has been little empirical research on category structure in the domain of conceptual modeling. In general the field of conceptual modeling lacks empirical research that tests (cf. [7, 20, 19]), while in this particular case work on formalizations and tools to support graded structures has already been done (e.g, [27, 6]). The focus of this work is thus to present an exploratory *empirical* investigation into the structure of categories from modeling languages to determine whether the potential issues we described realistically come into play (i.e., *there are categories from modeling languages that are of a graded nature*). Based on our findings we will discuss how an understanding of these categories can be used to guide the process of model creation and use, for instance by helping modelers and stakeholders in capturing as much useful information about the allowed range of instantiation for a model, enabling others to read and use the model as it was intended by the creators.

The primary findings that we will show in this paper are that most of the categories from modeling languages tend to exhibit a graded structure, that many of the terms used for them are considered partial members, while a surprising amount of terms are also considered clear non-members. The possible complications that might arise because of these and other findings, and what kinds of models they affect are discussed in more detail in the rest of the paper.

The remainder of this paper is structured as follows. We detail our experimental setup in section 2, present the results in section 3, and discuss the consequences they may have for modeling and modeling languages in section 4. Finally, in section 5 we conclude and propose directions for future research.

2 Experimental setup

What we wish to achieve is examine whether a number of categories more closely resemble graded or discrete categories. We can do this by performing a category membership experiment for the target categories and a number of benchmark

categories of which we know whether they are typically judged in a discrete or graded fashion, and to what extent their members are judged so.

2.1 Considerations

There are a number of considerations to take into account with this investigation. First is the issue of the potential participants and their (natural) language. Most importantly, when we ask whether a certain thing is a member of a category or not, we would optimally do that in the participant's native language. However, as the terms used by most modeling languages and methods (i.e., the terms we will use in our experiment) are in English, we need to either use them as-is, or translate them. Given that most modelers use the terms as given by languages (i.e., in English), albeit often appending their own semantics, we will perform the experiment with the terms without localizing them.

For the benchmark we will use datasets from previous research. However, an issue with the existing and still often used datasets is that they can be outdated (e.g., the commonly used Barr & Caplan dataset was published in 1987), and they can be sensitive to cultural differences. Category judgments can shift as certain objects fall out of common use and are replaced by entirely different things, as well as certain objects can be seen differently in different cultures. For example, while in Barr & Caplan's dataset bicycles are found to not be strong members of the category VEHICLE, repeating the experiment with Dutch, Danish or German participants (who are far more likely to use a bicycle as a mode of transport [21]) will likely lead to significantly different results. As such, care will have to be taken when interpreting the results from the benchmark categories to place them into the correct frame of time and culture. While there are other datasets available that were gathered from non-English native speakers (e.g., Ruts et al. [24] who performed an exemplar generation study amongst Belgian students) that might be used to create a more even dataset, they often only include full members and lack the necessary borderline and non-members.

Finally, there is the question of the granularity of the categories from the modeling languages that we will investigate. On the highest level there is the distinction between entities and relationships (and sometimes values), which are the main categories used by certain non-domain-specific languages (e.g., ER, ORM). However, it would be more interesting to look into the more specific categories (e.g., PROCESS, RESOURCE, ACTOR) used by domain-specific languages (e.g., BPMN, e3Value, ArchiMate) as they are more likely to yield discriminating results. This will also make it possible to eventually distinguish between groups with different focuses (e.g., the BPM community, the ArchiMate community) and find out if there are significant differences between them in terms of categorization. Thus, for this investigation we will focus on categories found in domain-specific languages.

2.2 Method

Participants: Fifty-six participants participated in the present study. Twenty-one of them were advanced (3^{rd} or 4^{th} year) students at an undergraduate

university of applied science with a focus on computing science and modeling, thirty-five were professional modelers employed at a research institute with a focus on IT and used modeling languages and tools to varying degrees. All participated voluntarily and received no compensation for their participation.

Materials: The materials used for the benchmark in the experiment were based on the list of exemplars reported on by Barr & Caplan [4]. We used 5 full, 5 partial and 5 non-members terms for both of the benchmarks. They were translated and presented in Dutch for the twenty students, but presented in English for the participants at the public research center, given that this was the only shared language between all participants and all participants were sufficiently fluent. In this text we consistently refer to them in English. For this benchmark we included the categories FRUIT and VEHICLES (see Table 4 in the appendix). For the modeling part of the experiment we investigated the categories ACTOR, EVENT, GOAL, PROCESS, RESOURCE, RESTRICTION and RESULT. These categories and related terms result from an earlier performed analysis on modeling languages and methods commonly used in enterprise modeling, which was reported on in [18]. The terms used for the members of these categories are the terms as used by the modeling languages and methods, based on the official (or most-used) specification (see Table 2 in Ref. [18] for the entire list, not replicated here due to space considerations).

Procedure: The procedure was based on Estes' [9] setup. Participants were divided into three groups (students, beginning modelers and expert modelers) and completed the task through on online survey. In this survey, participants were instructed to judge whether a list of given terms were either full, partial or non-members for the current category. Participants were informed beforehand that partial member scores meant that the exemplar belonged to the category, but to a less degree than others. This was first done for the two benchmark categories, and followed in the same way for each of the investigated categories from the modeling languages. The orders of the terms in each category were randomized for each participant. Care was taken to validate that participants filled out the survey seriously by comparing results and checking for long strings of repeating answers that the randomization should have prevented from occurring.

3 Results

The proportion of graded membership judgments for the terms used in the benchmark which are partial members are shown in detail in Table 1. The terms listed here are solely the partial members (as determined by the original datasets). What was to be expected is that the typically discrete category (FRUIT) would show lower proportions of graded judgments compared to the typically graded category (VEHICLES). The given scores indicate the proportion of partial member judgments (e.g., 19% of students, 13% of beginning modelers, and 30% of expert modelers considered an avocado as a partial member of the FRUIT cate-

gory). Shown are respectively the scores for students, beginning modelers, expert modelers, and the scores as reported by Barr & Caplan [4], and Estes [9].

Table 1: Partial member proportions for the partial member terms of the benchmark.

Category	Term	Student	Beginner	Expert	Ref. [4]	Ref. [9]
FRUIT	avocado	0.19	0.13	0.30	0.37	0.16
	coconut	0.24	–	0.05	0.38	0.37
	tomato	0.33	0.27	0.25	0.34	0.05
	cucumber	0.19	–	0.25	0.23	0.21
	rhubarb	0.14	0.20	0.15	0.45	0.26
VEHICLES	gondola	0.24	0.20	0.20	0.50	0.21
	tricycle	0.14	0.13	0.10	0.64	0.58
	wheelchair	0.29	0.27	0.50	0.70	0.63
	horse	0.48	0.27	0.55	0.54	0.50
	husky	0.38	0.27	0.55	0.27	0.21

A more detailed overview of the average amount of full, partial and non-member judgments for each investigated category is given in Table 2. The results are given for each investigated group (students, beginning modelers and expert modelers), and indicate the proportion of membership judgments. For example, students considered 47% of the presented terms for the ACTOR category to be full members, 18% to be partial members and 35% to be non-members. The primary points of interest here are the higher scoring partial and non-member results, as they indicate words actually used by modeling languages that are either only considered to be partially reflective of their category (e.g., a 'market segment' would be only considered somewhat an ACTOR), or are considered not to be exemplars of that category (e.g., a 'requirement unit' would not be considered an ACTOR).

Table 2: Average amount of membership scores (full, partial and non-members) for each group of investigated categories.

	student ($n = 20$)			beginner ($n = 15$)			expert ($n = 21$)		
Category	full	partial	non	full	partial	non	full	partial	non
ACTOR	0.47	0.18	0.35	0.30	0.14	0.55	0.41	0.25	0.35
EVENT	0.46	0.14	0.41	0.39	0.16	0.45	0.29	0.19	0.51
GOAL	0.65	0.11	0.23	0.60	0.16	0.24	0.56	0.20	0.24
PROCESS	0.66	0.14	0.20	0.62	0.22	0.16	0.41	0.32	0.28
RESOURCE	0.59	0.19	0.22	0.62	0.19	0.20	0.54	0.22	0.24
RESTRICTION	0.50	0.21	0.29	0.55	0.18	0.27	0.39	0.24	0.37
RESULT	0.73	0.16	0.11	0.86	0.07	0.08	0.76	0.16	0.09
FRUIT	0.44	0.10	0.45	0.47	0.05	0.42	0.49	0.09	0.41
VEHICLE	0.48	0.14	0.37	0.49	0.13	0.37	0.51	0.20	0.29

Table 3 gives a detailed overview of specific modeling language terms considered partial members by at least $\geq 30\%$ of one of the investigated groups. A clear difference can be seen between the groups for most categories, with the expert modelers displaying on average a much higher amount of graded judgments than the students or beginning modelers. On average students considered 15% of the investigated terms to be partial members, while beginning modelers did so for 32% and expert modelers considered 83% to be partial members.

Table 3: Terms considered partial members by $\geq 30\%$ of at least one group. The terms listed here are *only* those considered partial members, thus not including the terms considered full or non-members. The amount of terms listed here is respectively 43%, 32%, 26%, 48%, 50%, and 25% of the total amount of terms investigated for each respective category.

Category	Term	Student	Beginner	Expert
ACTOR	unit			✓
	requirement unit			✓
	infrastructural component	✓		✓
	organizational component			✓
	device			✓
	application software			✓
	organizational unit			✓
	hardware			✓
	software	✓		✓
EVENT	behavior			✓
	function			✓
	interaction			✓
	activity		✓	
	task		✓	✓
	service task			✓
	value activity	✓		✓
	contribution			✓
	operation			✓
GOAL	expectation	✓	✓	✓
	requirement			✓
	consumer needs			✓
	target			✓
PROCESS	organizational service			✓
	infrastructure service			✓
	information service			✓
	other service		✓	✓
	IT service		✓	✓
	service			✓
	sub flow		✓	✓
	process flow			✓
	dependency path		✓	
	game		✓	✓

Table 3: (cont.)

	task			✓
RESOURCE	artifact		✓	✓
	hd			✓
	location		✓	
	data object		✓	✓
	business object			✓
	object	✓	✓	
	data input			✓
	input			✓
	value object	✓		✓
	network device		✓	
	representation		✓	
	value port		✓	
	device	✓		
RESTRICTION	belief		✓	
	priority		✓	
	value		✓	
	interface	✓		
	catching			✓
	throwing	✓		✓
	license			✓
	trust			✓
	interrupting			✓
	non-interrupting			✓
	strategy			✓
	strategic objective	✓	✓	✓
RESULT	end event		✓	✓
	payoff			✓

4 Discussion

We will first discuss the results in general, showing how they support the assumption that there are categories in modeling languages that are of a graded nature. We will then discuss in more detail to what kind of models and modeling languages our results are most applicable and consequences our findings entail for them. Finally, we also discuss a number of limitations of our current study that should be kept in mind when interpreting the results.

4.1 General discussion

It was expected that the partial member judgments for the natural and artifactual benchmark categories would show a difference, with the artifactual category

displaying a higher proportion of graded judgments. Although compared to the results from Barr & Caplan [4] and Estes [9] the overall amount of graded judgments seems to be lower, the relative distribution still seems intact. This is the case for both the beginning and expert modelers (the proportion of some graded judgments for VEHICLES being at least twice as large compared to the ones for FRUITS). This is not the case for the student group, as the difference between the benchmark categories there was found to be much smaller. This could be explained by the lower amount of experience with (and exposure to) modeling (and modeling languages) students have. This is further reflected in Table 3 where there are far less words considered partial members by students than by the more experienced modelers.

On average the proportion of partial member judgments is 0.16 for students, 0.16 for beginning modelers, and 0.23 for expert modelers. When we compare these scores to the average proportion of partial member judgments for the discrete and graded benchmark categories in Table 2 (respectively 0.10 and 0.14 for the students, 0.05 and 0.13 for the beginning modelers and 0.09 and 0.20 for the expert modelers), we can see that for the two groups of modelers most scores shown for the categories from modeling languages more clearly reflect the graded benchmark category than the discrete one. Thus, as a careful first investigation we seem to have found support that most categories from modeling languages are of a graded nature. Given that the distribution of terms for these categories was not the same as the benchmark categories (i.e., the benchmark categories were made up of equal amounts of full, partial and non-members, while for the categories from the modeling languages we were unaware of this distribution, with them likely containing proportionally more full members) this makes it all the more acceptable to support the idea described in the introduction that *these categories can be seen as exhibiting a graded structure.*

Another interesting finding is the high amount of non-member judgments found in many of the categories. It is striking that the terms we have used in modeling languages and methods are sometimes considered absolute non-members of their related category. In particular, it can be seen that EVENTS are the largest category for non-members across all groups (respectively $0.41, 0.45$, and 0.51), while ACTORS and RESTRICTIONS also have a high amount of non-members in some groups. A possible explanation for this is that people are quicker to judge about things they are specialized in, for example a process modeler having more snap judgments about concepts to do with processes, and thus also being more willing to rule out terms. In practice this means that the terminology we use originating from some languages might not reflect our innate category judgments at all, raising the question whether this is a bad thing (e.g., because the terminology is far away from our naive understanding and semantics) or perhaps not that much of a problem (e.g., because the mismatch between a term and our understanding of it in a given context makes it easier to 'redefine' and use it in that context).

As already hinted at and most clearly visible in Table 3, there is a striking difference between the groups we investigated when it comes to the proportion of partial member judgments. The expert modelers have a far higher amount of

partial member judgments compared to the students, and in a lesser degree to the beginning modelers. An exception to this are PROCESSES and RESOURCES, which are judged more comparably between beginning and advanced modelers. This might be explained by the fact that the department of the research institute which the majority of the participants were working in has a strong focus on service science and is thus focused on many efforts involving processes (e.g., process modeling). An explanation for the difference between these groups might be that students simply have had less exposure to modeling terminology and are thus more likely to give absolute judgments. On the other hand, there is also the possibility that the (expert) modelers are, through training and experience, cognitively better equipped to deal with situations with abstract and vague concepts (cf. [26]), which could manifest in a higher amount of graded judgments.

4.2 Applicability of our findings

Before we move on to discuss the consequences of these findings for model creation and use, we need to specify more clearly to what kinds of models and languages they are applicable. Models created with more general modeling languages like UML, ER, and ORM are less affected by the existence of graded categories, as the main types (i.e., entities and relationships) they use are already so abstract that one would not so much expect *subtle* misunderstandings that stay unnoticed to arise in the same way as they would in domain specific languages. Furthermore, when languages like these need to be made more specific, they can do so by, e.g., explicitly capturing the necessary facts in ORM, or using UML stereotyping to create the needed new semantics. The semantics given by the modelers can then become an explicit part of the language.

However, when it comes to domain-specific languages our findings become much more relevant. This is because the semantics of the types used by (and often pre-defined in) these languages are less abstract than the ones mentioned above, and the risk of subtle misunderstandings that are not immediately noticed is higher. With the plethora of domain-specific languages (e.g., ArchiMate, BPMN, e3Value, i*, ITML, ADeL) in active use today all with their own focus (e.g., enterprise architecture, processes, value exchanges, goals, IS implementations, IS deployments) our findings could have consequences for many modeling efforts. The consequences we discuss should thus be taken to be most relevant for domain-specific modeling languages like these and any artifacts based on the models created with them.

4.3 Consequences for modeling

When it comes to the modeling languages and models that are affected by our findings, we see a number of different kinds of models:

1. *models used to communicate* between, and with different modelers and stakeholders (e.g., conceptual models)
2. *models used to formalize* information from a given domain, for whatever purpose (e.g., ontologies, models as documentation)

3. *models used to execute* by non-human systems (e.g., compiled source code)

This list is not intended as a taxonomy of models, nor as an exhaustive list of the different kinds of models that are affected by graded category structures. It is merely a starting point to reason about the different consequences we see our work having for different kinds of models. We furthermore do not mean to imply that kinds of models are mutually exclusive (e.g., that models used to communicate are never used to formalize or transformed into executable models).

Models used to communicate involve conceptual models of many possible purposes (e.g., capturing a domain, models used to guide decision making). As we have shown that the categories used by modeling languages are likely of a graded nature, the models created by them necessarily also contain categories of a graded nature. The most important consequence here is that an instantiation of a model is not just simply valid or invalid, but will display degrees of validity as well. If the category goal is seen as a graded structure, with some things being better goals than others, it is thus possible to instantiate a model that contains some goal type with two different cases that are both valid, but not equally so. As the formal semantics of most modeling languages do not explicitly support such degrees of validity, it is important that we are clear about the limits of conceptual validity of our models. In other words, to ensure people read and use models in a similar way, we need to ensure that we provide clear examples of possible valid instantiations, and perhaps more importantly, clear examples of that which we consider invalid as well.

For example, while 'hardware' and 'software' are both considered partial members (by at least the experienced modelers in our study) of the category ACTOR, the exact degree to which they are both considered so is something that is likely different for different (groups of) people. If we are creating a model used for the implementation of an information system, which would likely incorporate such terms for the things that act to support and execute business activities, we need to be clear to what degree they can both be seen as ACTORS. For instance, the modelers or stakeholders might envision the hardware as the actual acting part, with the software providing the instructions for doing it so, and thus find a model where 'hardware' is said to act out a business function more valid than where 'software' does so. However, others might disagree and see 'software' as the actual thing that acts. As these interpretations can be different from group to group, it is thus important to involve explicit discussions about the degrees of validity for different things we use in our models during model creation.

Models used to formalize are for instance models that capture knowledge about a certain domain and attempt to formalize it in order to reduce the amount of ambiguity. A formalization involving graded categories needs to ensure that membership requirements are not discrete, and more important, take into account the relevant properties of a graded category (e.g., centrality and typicality of members). There is work in the field of ontology engineering that strives towards explicitly supporting these structures, e.g., [2] and explicit modification of ontology formalizations to incorporate the noted features [27, 28], and critiques and extensions of proposed work, e.g., [6]. If such formalizations are not used, and

instead a classical approach based on discrete judgments is used, much semantic information about the domain and the judgments from the original modelers is lost. This can lead to misinterpretations by other readers and users of the model if there is no communication between them and the original modelers anymore. For instance, someone might consider a horse as a VEHICLE (albeit an atypical one) and thus consider it to be somewhat of a valid vehicle in their created ontology. However, when this is formalized discretely, any other member of the VEHICLE category (e.g., a car) would be considered on equal footing with the horse, while this has no grounding in the real world whatsoever. As such, the formalization can no longer be considered a correct representation of the real world and loses a lot of its value.

Models used to execute are for instance source code which is run by an interpreter, or compiled and then executed. Other options are models interpreted by model provers, expert systems, or ontologies used for automated reasoning and so on. For example, a model used by an expert system to check for a number of possible cases (e.g., a medical advice system) might need graded structures and judgments in order to correctly reason with the real-world information. A number of formalizations for e.g., descriptive logics have been proposed to incorporate graded features like typicality and centrality [5, 12, 11]. These models are affected in a similar way to the ones used to formalize, meaning that their formalizations need to support any graded structures found in them. This is all the more important to ensure here, as executable models are often no longer read and interpreted by people, and thus any errors or oversights in them are less likely to be corrected.

4.4 Limitations

While it is good to find that our results hint towards the modeling categories having a graded nature, care must be taken not to immediately extrapolate this finding and use it to judge the structure of the investigated modeling categories in general. For one, this has been only one study, with two of our groups of participants being people with *professional* experience in conceptual modeling. For these reasons repeating the study presented here with additional groups of (experienced) people to validate whether they share the same graded structure would be a prudent thing to do.

Furthermore, as categorization judgments are something inherent to people, it would also be useful to perform this study on specific subgroups of modelers (e.g., process modelers, enterprise architects, goal modelers) to analyze whether the proportion of graded responses is different for specific categories or not (i.e., test whether categories that modelers are focused on receive less partial member judgments). One could for instance hypothesize that people who are specialized in a topic have less semantic flexibility in regards to the categories of that topic.

Related to the terms we used, it might also be interesting to see whether the introduction of model context (i.e., presenting the terms while being used in a model) instead of the isolated terms themselves would yield different results. Nonetheless, the results from our study investigating the terms in isolation also

provides useful insight into the amount of terms that would typically not be considered as good representatives of their functional category. Furthermore, this might provide an additional source of complexity and confusion for participants, as with the amount of terms we used in the study, a large amount of different modeling languages would be used, some of which participants are likely not familiar with.

It should also be noted that the study presented here talks about the structure of the category in terms of it being graded or discrete, but does not aim to give a representation of the *internal* structure. Further studies involving explicitly eliciting typicality and centrality of the terms investigated here could be done in an attempt to discover such structures. It is very likely that the internal structure of the categories (which is regardless of the graded or discrete question) is specific to different groups of people, as it can be readily expected that process modelers will have a different central core for a number of categories than, for example, goal modelers. Thus, such studies should also be performed with a number of different groups of modelers.

Finally, as referred to earlier, the distribution of the terms for the modeling categories was not optimal (i.e., not evenly divided between full, partial and non-member), which makes it more difficult to infer detailed general statements about the structure. Such work on the detailed structure of these categories like described above can be undertaken in further research, where the individual category members are rated on typicality and centrality in order to attempt to build an actual representation of a shared category structure. Such findings could then be used to create a more evenly distributed set of modeling terms for further membership judgment experiments.

5 Conclusion

We have presented a study into the category structure of types used by most modeling languages. This study showed that many of these modeling categories are likely of a graded nature (that is, some things are considered to be better members than others), which can have an effect on the semantics of models and their derivatives. We have discussed the implications for validity of models and proposed that more study into the understandings specific groups have of such categories would be a worthwhile avenue of research. The main contribution of this work has been empirically showing that *the categories we use to model are likely of a graded nature*, which before was only assumed (or worse, ignored). More specifically, we have shown that the modeling terminology from actual modeling languages and methods are affected by this graded nature as well. In future work we hope to extend this research to different groups with a strong focus on a specific domain to investigate potential categorization differences between different people operating in different domains.

Acknowledgements. This work has been partially sponsored by the *Fonds National de la Recherche Luxembourg* (`www.fnr.lu`), via the PEARL programme.

References

1. Adelson, B.: Comparing natural and abstract categories: A case study from computer science. Cognitive Science **9**(4), 417 – 430 (1985)
2. Aime, X., Furst, F., Kuntz, P., Trichet, F.: Conceptual and Lexical Prototypicality Gradients Dedicated to Ontology Personalisation. In: R. Meersman, Z. Tari (eds.) On the Move to Meaningful Internet Systems: OTM 2008, *Lecture Notes in Computer Science*, vol. 5332, pp. 1423–1439. Springer Berlin / Heidelberg (2008)
3. Altarriba, J., Bauer, L.M.: The distinctiveness of emotion concepts: A comparison between emotion, abstract, and concrete words. The American journal of psychology pp. 389–410 (2004)
4. Barr, R., Caplan, L.: Category representations and their implications for category structure. Memory & Cognition **15**(5), 397–418 (1987)
5. Britz, K., Heidema, J., Meyer, T.: Modelling object typicality in description logics. In: A. Nicholson, X. Li (eds.) AI 2009: Advances in Artificial Intelligence, *LNCS*, vol. 5866, pp. 506–516. Springer Berlin (2009)
6. Cai, Y., Leung, H.f.: A formal model of fuzzy ontology with property hierarchy and object membership. In: Li et al. (ed.) ER, *LNCS*, vol. 5231, pp. 69–82. Springer Berlin (2008)
7. Davies, I., Green, P., Rosemann, M., Indulska, M., Gallo, S.: How do practitioners use conceptual modeling in practice? Data & Knowledge Engineering **58**(3), 358 – 380 (2006)
8. Diesendruck, G., Gelman, S.: Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. Psychonomic bulletin & review **6**(2), 338–346 (1999)
9. Estes, Z.: Domain differences in the structure of artifactual and natural categories. Memory & cognition **31**(2), 199–214 (2003)
10. Estes, Z.: Confidence and gradedness in semantic categorization: Definitely somewhat artifactual, maybe absolutely natural. Psychonomic bulletin & review **11**(6), 1041–1047 (2004)
11. Freund, M., Descles, J.P., Pascu, A., Cardot, J.: Typicality, contextual inferences and object determination logic. In: FLAIRS, vol. 4, pp. 491–495 (2004)
12. Giordano, L., Gliozzi, V., Olivetti, N., Pozzato, G.: Reasoning about typicality in preferential description logics. In: Hoelldobler et al. (ed.) Logics in Artificial Intelligence, *LNCS*, vol. 5293, pp. 192–205. Springer Berlin Heidelberg (2008)
13. Hampton, J.A.: Similarity-based categorization and fuzziness of natural categories. Cognition **65**(2-3), 137 – 165 (1998)
14. Hampton, J.A., Dubois, D., Yeh, W.: Effects of classification context on categorization in natural categories. Memory & Cognition **34**(7), 1431–1443 (2006)
15. Henderson-Sellers, B.: UML - the Good, the Bad or the Ugly? Perspectives from a panel of experts. Software and System Modeling **4**(1), 4–13 (2005)
16. Hoppenbrouwers, S.J.B.A.: Freezing language : conceptualisation processes across ict-supported organisations. Ph.D. thesis, Radboud University Nijmegen (2003)
17. Kalish, C.W.: Essentialism and graded membership in animal and artifact categories. Memory & Cognition **23**(3), 335–353 (1995)
18. van der Linden, D.J.T., Hoppenbrouwers, S.J.B.A., Lartseva, A., Proper, H.A.: Towards an investigation of the conceptual landscape of enterprise architecture. In: T. Halpin et al. (ed.) Enterprise, Business-Process and Information Systems Modeling, *LNCS*, vol. 81, pp. 526–535. Springer Berlin Heidelberg (2011)

19. Moody, D.: Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. Data & Knowledge Engineering **55**(3), 243–276 (2005)
20. Persson, A., Stirna, J.: Why enterprise modelling? an explorative study into current practice. In: Dittrich et al. (ed.) Advanced Information Systems Engineering, *LNCS*, vol. 2068, pp. 465–468. Springer Berlin (2001)
21. Pucher, J., Buehler, R.: Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. Transport Reviews **28**(4), 495–528 (2008)
22. Rosch, E.: Natural categories. Cognitive Psychology **4**(3), 328–350 (1973)
23. Rosch, E., Mervis, C.B.: Family resemblances: Studies in the internal structure of categories. Cognitive Psychology **7**(4), 573 – 605 (1975)
24. 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 **36**, 506–515 (2004). 10.3758/BF03195597
25. Sowa, J.: The Role of Logic and Ontology in Language and Reasoning. In: Theory and Applications of Ontology: Philosophical Perspectives, pp. 231–263. Springer (2010)
26. Wilmont, I., Barendsen, E., Hoppenbrouwers, S., Hengeveld, S.: Abstract reasoning in collaborative modeling. In: HICSS, pp. 170–179 (2012)
27. Yeung, C.A., Leung, H.F.: Ontology with likeliness and typicality of objects in concepts. In: ER, pp. 98–111 (2006)
28. Yeung, C.A., Leung, H.F.: A formal model of ontology for handling fuzzy membership and typicality of instances. Comput. J. **53**(3), 316–341 (2010)

Appendix

Table 4: The categories and terms for the benchmark as adapted from [4] and [9], followed by the used Dutch translations for the student group.

Category	Term
FRUIT (discrete)	apple, pear, plum, banana, pineapple, avocado, coconut, tomato, cucumber, rhubarb, carrot, onion, potato, rose, spinach
VEHICLES (graded)	bus, car, truck, van, taxi, gondola, tricycle, wheelchair, horse, roller skates, husky (dog), lawnmower, bus driver, carton, newspaper
FRUIT (discrete)	appel, peer, pruim, banaan, ananas, avocado, kokosnoot, tomaat, komkommer, rabarber, wortel, ui, aardappel, roos, spinazie
VEHICLES (graded)	bus, auto, vrachtwagen, busje, taxi, gondel, driewieler, rolstoel, paard, rolschaatsen, husky (hond), grasmaaier, buschauffeur, doos, krant