Empirically evaluating modeling language ontologies: the Peira framework

Sotirios Liaskos[1*], Saba Zarbaf[2], John Mylopoulos[3], Shakil M. Khan[4]

[1*]School of Information Technology, York University, 4700 Keele St., Toronto, M3J 1P3, Ontario, Canada.
[2]Department of Electrical Engineering and Computer Science, York University, 4700 Keele St., Toronto, M3J 1P3, Ontario, Canada.
[3]Department of Computer Science, University of Toronto, 40 St.George St., Toronto, M5S 2E4, Ontario, Canada.
[4]Department of Computer Science, University of Regina, 3737 Wascana Parkway, Regina, S4S 0A2, Saskatchewan, Canada.

*Corresponding author(s). E-mail(s): liaskos@yorku.ca;
Contributing authors: sabazb@cse.yorku.ca; jm@cs.toronto.edu; shakil.khan@uregina.ca;

Abstract

Conceptual modeling plays a central role in planning, designing, developing and maintaining software-intensive systems. One of the goals of conceptual modeling is to enable clear communication among stakeholders involved in said activities. To achieve effective communication, conceptual models must be understood by different people in the same way. To support such shared understanding, conceptual modeling languages are defined, which introduce rules and constraints on how individual models can be built and how they are to be understood. A key component of a modeling language is an ontology, i.e., a set of concepts that modelers must use to describe world phenomena. Once the concepts are chosen, a visual and/or textual vocabulary is adopted for representing the concepts. However, the choices both of the concepts and of the vocabulary used to represent them may affect the quality of the language under consideration: some choices may promote shared understanding better than other choices. To allow evaluation and comparison of alternative choices, we present Peira, a framework for empirically measuring the domain and comprehensibility appropriateness of conceptual modeling language ontologies. Given a language ontology to be evaluated, the framework is based on observing how prospective language users classify domain content under the concepts put forth by said ontology. A set of metrics is then used to analyze the observations and identify and characterize possible issues that the choice of concepts or the way they are represented may have. The metrics are abstract in that they can be operationalized into concrete implementations tailored to specific data collection instruments or study objectives. We evaluate the framework by applying it to compare an existing language against an artificial one that is manufactured to exhibit specific issues. We then test if the metrics indeed detect these issues. We find that the framework does offer the expected indications, but that it also requires good understanding of the metrics prior to committing to interpretations of the observations.

Keywords: Conceptual Modeling, Modeling Language Quality, Empirical Methods, Peira

1 Introduction

Developing conceptual models is an essential activity for the effective planning, development and maintenance of software-intensive systems [1–5]. Conceptual models (henceforth simply: models) facilitate communication among stakeholders, allow transformations from system requirements to architecture, design, and code, and, through formalization and automated reasoning, support validation and decision making.

A fundamental requirement for a useful model is that its stakeholders have a shared understanding of what the model means. Towards this end, (conceptual) *modeling languages* have been proposed, which offer guidelines to modelers on how to build models, and on stakeholders on how to read and interpret these models [6]. At the core of every modeling language lies an *ontology* that captures a shared *conceptualization* that language users (modelers and model readers) have of a given domain [7]. Thus, languages for discrete processes, adopt ontologies containing concepts such as *state*, *transition*, and *guard condition* (e.g., UML [8]) while languages for representing stakeholder intentions for requirements engineering use concepts such as *goal*, *task* and *actor* (e.g., iStar [9]). When designing a language, designers must decide what ontology the language should offer and with what natural language terms (e.g., "state", "goal", etc.) or other symbols (e.g., shapes, lines, etc.) these concepts should be referred to. Such choices are fundamental for a shared understanding of models.

Given a domain, one ontology may be a better choice than another with respect to bringing about a shared understanding. Moreover, for any concept in a chosen ontology, one choice of a term may describe the concept better than another. Traditionally, such choices seem to have largely been the prerogative of language designers based on their experience in the domain and in language design. There is, however, evidence that original decisions of language designers can be suboptimal and warrant future updates based on user feedback. For example, since its inception, a language for enterprise architecture (EA) modeling, Archimate [10], underwent a series of updates from version 1.0 (2009) to versions 2.0 (2012), 3.0 (2016) and 3.2 (2023). At each stage, its ontology was revised and enriched based on considerations that included, reportedly, user feedback [11–13]. As another example, a language for modeling stakeholder intentions, *i**, originally introduced in the mid-nineties and intensively studied for two decades thereafter also evolved into a new version, iStar 2.0 [9] . The latter was the outcome of a systematic consultation process involving feedback from researchers/users of the language and involved updates in the original ontology – such as, for example, replacement of the *soft-goal* concept with the *quality* concept, citing inconsistent use of the former in the community. While these changes can in part be the result of evolution of the target domains (EA, for example) or the community's thinking about how these domains should be modeled, a large part appears to simply be correction of a misalignment between the original ideas of the designers of the language and the needs of the audience of the language.

In light of such potential misalignments, language designers could benefit from integrating in the design process the collection of empirical evidence on how the intended language users understand and use a proposed set of concepts when performing modeling tasks. Such an evidence-based approach becomes more accessible, reliable, reproducible, and cost-effective when systematic ways are available for supporting it, while also offering ways to analyze observations into metrics that are directly interpretable to specific language design issues and recommendations.

This paper proposes and evaluates Peira[1], a framework for the experimental evaluation of modeling language ontologies. The framework is based on the adoption and extension of a system of ontology qualities and the introduction of a set of metrics for measuring these qualities. The metrics are applied on data collected from experiments where prospective language users classify natural language descriptions of the domain under the concepts of the ontology in question. Moreover, the metrics measure in-between agreement among experiment participants as well as agreement with a gold standard representing the designers' authoritative classifications. Depending on the results of the evaluation, the ontology may be improved with respect to either or both the

[1] πείρα (/'pi.ra/), the Greek word for experience, trial, experiment.

choice of its concepts and/or the choice of terms or other visual signifiers to refer to these concepts.

To evaluate the proposed framework, we conduct an experiment where we consider the ontologies of two languages: the first one is adopted from iStar [9] and the second is a revised version of the first where specific alignment issues are deliberately introduced. The main goal of the experiment is to observe if the issues embedded in the second language can be detected by the proposed metrics. In addition, we aim at exploring if attitudinal data collected from the participants, in which they themselves grade the appropriateness of each of the ontologies under evaluation, correlate with the observations on how the participants actually decided to associate descriptions to concepts.

This paper extends our earlier work [14] by offering refined abstract metric definitions, a new experimental evaluation that is based on real rather than simulated data, an alternative type of instrument, experimental treatment of relationships, and instrument reliability testing.

The rest of the paper is organized as follows. In Section 2, we present the key ideas and concepts pertaining to modeling language ontologies and their empirical evaluation. Building on these ideas, in Section 3, we present our system of metrics. Then, in Section 4, we present our experimental design, including the research questions, metrics, and hypotheses we investigate. In Section 5, we present the results of the study, followed by a discussion on validity threats and study conclusions in Section 6. Finally, in Section 7, we present related work and we offer concluding remarks and opportunities for future work in Section 8.

2 Background

2.1 Languages, conceptualizations, and vocabularies

One of the primary properties of an effective model is that everyone understands it in the same way. To support the development of models that evoke a common understanding among those who develop and use them, modeling languages have been proposed since the very early days of Software Engineering (SE) [15].

At the core of a modeling language lies an *ontology*, i.e., a specification of a conceptualization in the domain of interest. To more concretely develop the notion of an ontology, a concept, and a conceptualization we follow the Guarino et al. formulation of such in the context of a set of distinguished elements D, a.k.a. the Universe of Discourse (UoD), from a system S [7].

Firstly, an *extensional relation* (or simply *extension*) is a set of ordered n-tuples constructed with elements from D. Concepts represent the identification of such extensional relations under different states of the system S, i.e., under different worlds $w \in W$. Specifically an n-ary **concept** is a total function $\rho^n : W \mapsto 2^{D^n}$ from worlds to all possible n-ary relations on D. Further, a **conceptualization** is a set of concepts $\mathcal{R}$ on the domain space $< D, W >$.

For example, consider the system S to be a printer and D the set of aspects about the printer that we wish to talk about, such as, e.g., $D = \{$(isOn), (isOff), (powerButtonPressed), $\ldots\}$. The concept *trigger* maps possible, e.g., versions of the system to possible extensions over D. Hence in a world w_1, $\rho^1_{trigger}(w_1) = \{$(powerButtonPressed), (cancelButtonPressed), $\ldots$ $\}$; the right-hand side being the extension of *trigger* under w_1. In a world w_2, say, after the printer's firmware is updated, a different extension is mapped to the concept *trigger* that, e.g., includes (jamDetected). A conceptualization is a set of such concepts adopted by the modeling language. For example, in a language for state transitions a conceptualization includes $\{$*state, trigger, state transition, guard condition,* $\ldots\}$.

Concepts and conceptualizations need to somehow be represented to allow for communication among humans. A **vocabulary** V is hence constructed consisting of *signifiers*, each representing a concept of interest. In the simplest case, the signifiers are *terms*, i.e., words or phrases in a natural language. In the printer example, we might be interested in a vocabulary that contains terms such as "state", "trigger" and "state transition" to represent the concepts *state*, *trigger* and *state transition*. Indeed this is part of the vocabulary used by UML for StateMachine diagrams [8]. Different terms from potentially different natural languages could have been used to represent the same concepts. Likewise, languages may employ methods of visual signification of concepts instead of or in addition to having an explicit natural language construct. For example a box

inside a larger box indicates a containment relationship which may or may not have a name in the language definition. Although we have evaluated the proposed framework with linguistic terms only, the ideas and constructions we propose are generalizable to any kind of concept signification. For simplicity, we will henceforth use *term* and *signifier* interchangeably.

2.2 Ontological commitments and their sharedness

Once the vocabulary V is defined, it is important to ensure that the modeling *language* L in which it is embedded accepts models in accordance to the conceptualization. Firstly, a *model* for L, given D and a set R of n-tuples thereof is a total function $I : V \mapsto D \cup R$, mapping each vocabulary symbol $v \in V$ to an extension, i.e., a set elements or n-tuples from D. Obviously, we want the language to allow for each term $v \in V$ to map only to elements that are meaningful with respect to the concept that the term has been chosen to represent.

An *ontological commitment* represents exactly that. Formally, an ontological commitment is a mapping $\mathcal{I} : V \mapsto D \cup \mathcal{R}$, where $\mathcal{R}$ is a set of concepts, as defined above, which we wish to represent using the vocabulary V. In other words, an ontological commitment assigns meaning to signifiers/terms, thereby restricting the kinds of phenomena these signifiers/terms can represent.

Consider the part of UML used for the production of state diagrams. UML introduces the concepts *state* and *event*, and represents them with the vocabulary terms "state" and "event" and the corresponding visual signifiers; ovals and annotations on top of transition links. For a domain $D = \{$(isOn), (isOff), (powerButtonPressed)$\}$, only the first two elements can be in the extension of term "state" if we are to abide by the ontological commitment of the term to the concept *state*. Likewise for the same domain, only the last element (powerButtonPressed) is allowed to be in the extension of "trigger" if we are again to remain faithful to the ontological commitment of said term to the concept *trigger*.

The language designers' goal is that the ontological commitment of a language is *shared*, meaning that all or most users of the language associate signifiers to concepts – and consequently to extensions – in the exact way that the language designers intended them to. This relies on the signifiers evoking the right concept based on the user's previous understanding of the signifier (e.g., through use in natural language and daily life) and, potentially, but not necessarily, the user's prior study of the accompanying definitions and examples. In our example, UML designers interested in representing the concept *state* for English-speaking users of the language sensibly used the term "state" rather than the terms "class", "structure" or "κατάσταση".

One way to promote the sharedness of an ontological commitment is through introducing axioms in the language, such as meaning postulates [7], that restrict the models of the language to subsets that reflect the intended meaning of the terms. However, depending on the language and domain at hand, such axiomatizations may not always be available or easy to produce. As such, the choice of appropriate signifiers is often the key in evoking a consistent ontological commitment among users; assuming also that the conceptualization itself is well chosen.

2.3 Conceptual modeling as classification of domain phenomena

Language designers may draw confidence from their experience or analytical arguments that the choice of concepts to include in the language and the terms or other signifiers used to represent the chosen concepts is optimal for a specific user audience, e.g., requirements analysts, software designers, business process analysts, etc. However, the ultimate judge of this are the users of the language themselves, i.e., producers (modelers) and consumers (readers) of models, who are meant to use the language to successfully communicate among themselves.

Focusing on modelers and looking deeper into how they describe reality through constructing models, we find that at the heart of the process lies the task of *classifying* real world phenomena under the concepts represented by the offered signifiers. In a requirements modeling context, for instance, modelers are likely analysts who read large amount of information, such as documentation, interview transcriptions, policy books, etc., and classify chunks of that information under the

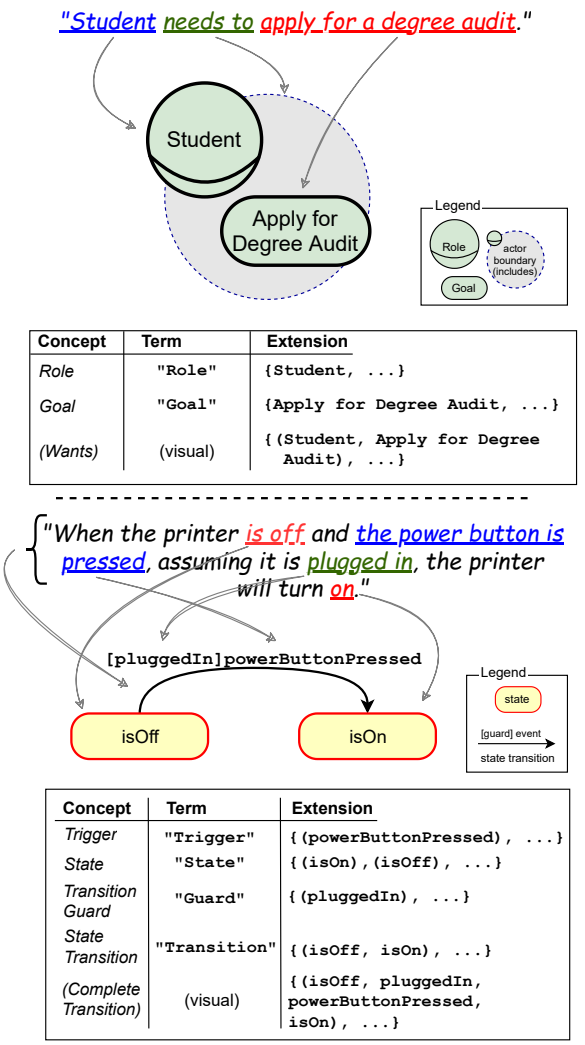

Figure 1 Model development as an extension formation process for an iStar 2.0 diagram [9] and a UML StateMachine diagram [8].

terms and/or visual signifiers that are available in the language.

Two examples are depicted in Figure 1. The upper part of the diagram depicts the translation of a chunk of domain information to part of a diagram drawn in iStar 2.0 [9] The lower part shows the same process for a UML StateMachine Diagram [8]. Thus, in iStar "role" instances are represented using a distinctive circular shape. When a modeler places such a circular shape in a goal diagram and writes "Student" in it, as is the case in the figure, she essentially classifies the domain element "Student" under the iStar language concept represented with the term "role". Likewise, the UML state diagram modeler classifies "powerButtonPressed" under "trigger" by placing it appropriately on the transition label. By further drawing and labeling the entire transition arrow she signifies that a quadruple formed by the four associated elements, i.e., the two states, the guard condition and the trigger, is classified under an unnamed signifier which would stand for a hypothetical StateMachine *complete transition* concept; i.e., a transition combined with a guard and a trigger.

Conversely, for the model reader, that a domain element has been classified under a specific term or signifier signals the desire of the modeler that the element is understood as an instance of the concept represented by the term, and that readers subsequently act (e.g., perform inferences, implement, validate, etc.) according to this information – a process that has been referred to as *activation* [16]. Hence, the model development practice is based on classifying chunks of domain information (representations of elements), while the act or reading a model is one that involves recognizing such classifications.

2.4 Ontology-vocabulary alignment and inter-rater agreement

As we saw, language designers desire that all users of the language understand it in the same way – preferably, the way the designers think is right. Thus, if we asked a number of different modelers to model the domain information *"the student needs to apply for a degree audit"* using iStar 2.0, a significant majority, including the designers, would hopefully perform the same classifications and produce the same or a very similar model. Reversely, by reading the model one would ideally be able to reproduce the same understanding of the domain as that of the modelers. Further, the designers would want to ideally agree with the way their terms are used by users. It follows that detection of disagreements in the way classifications are performed, firstly among the users of the language and then between users and designers constitutes a problem, as it defeats the purpose of clear communication.

We use the term *ontology-vocabulary alignment* to refer to the extent to which the signifiers in the vocabulary evoke understanding among users and

between users and designers that is indicative of such a shared ontological commitment. Detection of disagreement is an indication of misalignment – each user understands and adopts a different, if any, ontological commitment – which is, in turn, indicative of an issue with the choice of signifier or term to represent a concept, the way it is explained and/or exemplified, or, at a deeper level, with the choice of the concept it is meant to represent.

Agreement or disagreement among agents with respect to how they classify content to categories has been extensively studied in the context on qualitative content analysis [17]. In content analysis, units of content (e.g., text, images, audiovisual segments) taken from a domain are classified by raters under distinct categories (*codes*), so that the latter can then be used for the development of theories that are grounded on the domain information that was coded. Such grounding, however, can be considered reliable only if raters agree on how the content must be classified. Towards this end, various quantitative measures of inter-rater agreement have been proposed [18]. Absent or low observed inter-rater agreement may be due to, among other things, a suboptimal choice of codes.

The coding practices – and measurement of reliability thereof – employed in qualitative content analysis offer a model for us to follow for the measurement of vocabulary qualities. Thus, as with qualitative analysis, agreement among users on how signifiers are used to classify domain content is an indication of successful sharedness of the ontological commitment of the language.

2.5 Wand and Weber's misalignment characterization framework

While the wealth of inter-rater agreement metrics that have been introduced [18] can potentially be utilized as a first measure of ontology-vocabulary alignment, such general-purpose measures would not offer much detail with regards to the source and nature of possible misalignments. We can look for a more refined way of characterizing ontology-vocabulary misalignment in Wand and Weber's framework for comparing ontological with grammatical constructs [19]. In that framework the set of real-world constructs we want to represent (in our case: concepts) is distinguished from the set of grammatical constructs we use to represent the former (in our case: terms or other signifiers).

Two kinds of mappings between the two sets are proposed. The *representation mapping* is concerned with whether and how the concepts enjoy adequate and complete representation by the signifiers of the language. Conversely, the *interpretation mapping* is concerned with whether and how the signifiers that are put forth by the designers appropriately and completely correspond to concepts.

Ideally, both mappings are total and 1-1. From the representation mapping standpoint, every concept must somehow be represented by a signifier. If that is the case, we have *ontological completeness*, otherwise we have *construct deficit*. When we have construct deficit, there is at least one concept in the conceptualization that is not represented by any of the vocabulary terms. In addition, the representation mapping needs to be 1-1, meaning that each concept must be represented by exactly one signifier (*ontological clarity*). This is not satisfied when a signifier in the vocabulary represents more than one concept – so it is unclear which concept it represents each time it is used. Such phenomenon is called *construct overload* in the framework.

Focusing on the interpretation mapping we again are interested in the same properties. The mapping must be total in that every signifier must stand for a concept. Should there be one or more signifiers that do not clearly associate with any of the concepts included in the language, we have *construct excess*, i.e., the signifier may not be needed. Likewise, the mapping may not be 1-1, meaning that two or more different signifiers may be representations of the same concept. In that case we have *construct redundancy*.

The four categories of issues in the ontology-vocabulary alignment, can be useful for identifying, characterizing, and fixing issues with candidate language vocabularies. For example, if language designers are told that the language suffers from construct excess, they would know that the course of action for fixing the problem – if it is indeed a problem and not a deliberate property of the language – is to identify the superfluous term and consider removing it from the language.

To analyze the representation and interpretation mappings so as to detect such issues, presumes a way to observe how language users associate vocabulary terms, which are directly observable representations, with concepts, which are

unavailable to direct observation. However, observing how users classify domain elements under vocabulary terms (effectively, recall, constructing models I of the vocabulary) offers us a window into the ontological commitment $\mathcal{I}$ that the users are operating under, and, hence, implicitly, the concept they have associated the term with. Hence, by appropriately analyzing various ways by which users disagree about their classifications we can detect possible issues with the representation and interpretation mappings – excess, deficit, etc. – while also comparing the users' apparent ontological commitment with the one expected by the designers.

Peira offers a set of metrics that are designed to detect each of these classes of issues from sets of classification data. In the next section we describe these metrics in detail, along with all the other constituents of the Peira framework.

3 The Framework

3.1 Overview

The Peira framework consists of a set of measurement concepts that describe the logic and process of data collection as well as set of abstract metrics to be used for analyzing the corresponding data collected. These two components and their constituent concepts can be viewed in Figure 2. Application of the framework aims at systematically developing a *Data Set* of research participant *Ratings* and then utilizing a system of *Metrics* for translating the data into *Indications* of quality issues of the language under investigation. The metrics are distinguished in two categories. *Rater-authoritative* metrics compare the ratings of a sample of participants with normative ratings that represent the intent of the language designers, while *Within-rater metrics* do not assume the presence of such normative ratings but are based on different ways by which the participant ratings agree or disagree with each other. Below we discuss the measurement concepts, the within-rater, and the rater-authoritative metrics in sequence.

3.2 Measurement concepts

The proposed measurement framework requires the following components.

- A set of signifiers V, i.e., a vocabulary that needs to be evaluated.
- A set of human raters P who perform a number of rating tasks.
- A set of descriptions E taken from sample application domains.
- A set D of distinguished elements from the domain and a set R of n-tuples constructed using D. Let $\mathbf{D} = D \cup R$ for convenience.

All sets are defined by the designers of the vocabulary or its evaluators. The vocabulary V is the set of signifiers that a modeling language under evaluation uses to refer to its ontology. The raters $p \in P$ are samples taken from the population of the intended users of the language. The descriptions $e \in E$, offer natural language presentations and contextualizations of possible worlds, i.e., states of a system. The discourse elements and n-tuples thereof $d \in \mathbf{D}$ are extracted from the descriptions as items that the language designers or evaluators believe should be modeled by the user of the language using items from V. It is further assumed that $\mathbf{D}$ contains a sufficient number of representative instances of all concepts of interest and includes no elements that are not understood by the designers/evaluators to be instances of any concept of interest. The same discourse element may or may not be relevant in one or more descriptions from E. We use the term *subject* to refer to a domain element under a specific description. Hence, the set S of all subjects (d, e) is drawn from $\mathbf{D} \times E$.

As an example, consider a hypothetical goal-based enterprise modeling language that includes the terms *"agent"*, *"assessment"*, *"goal"*, *"intention"*, *"believes"*, *"plays-role"*, *"motivates"*; the example vocabulary is inspired by iStar 2.0 [9] and Archimate [10]. These seven terms constitute the vocabulary V to be evaluated. Seen as predicates, the first four are unary and the last three binary, i.e., represent entities and relationships, respectively. The designers of the language wish to evaluate V in its use for enterprise modeling. For the purpose, they produce descriptions of states of a real or fictional enterprise that contain facts and phenomena the language is supposed to capture. The specific facts and phenomena are then identified as discourse elements that need to be modeled.

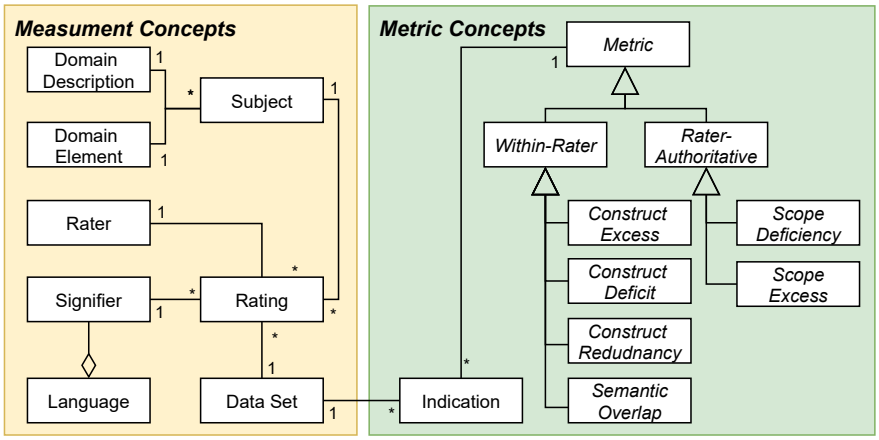

Figure 2 Peira's conceptual framework.

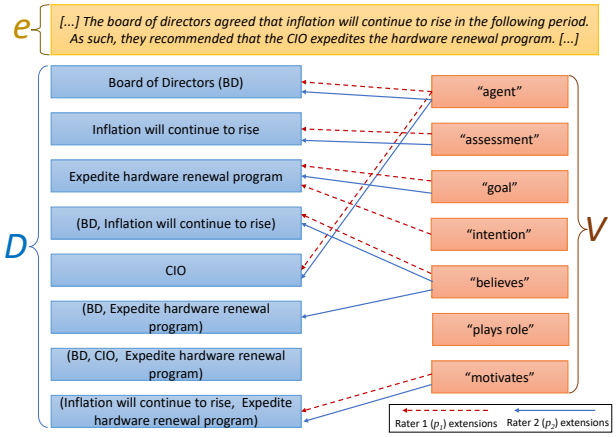

Figure 3 Rating example.

An example can be seen in Figure 3. From the description e (top of the figure), the evaluators have identified a set $\mathbf{D}$ of distinguished n-tuples of discourse elements (left side of the figure). The evaluators assume that this sample completely exemplifies the concepts that need to be included in the language, in a sense that every concept of interest is represented by a number of instances in the set. In the examples, the elements are unary, e.g., "CIO", "Expedite hardware renewal program", tuples, such as "(BD, Inflation will continue to rise)" or triples, such as "(BD, CIO, Expedite hardware renewal program)".

A sample of enterprise architects, who have just been introduced to the language, are then shown V, e and $\mathbf{D}$ and are asked to form the extension of each of the terms in V under e. In effect the raters are asked to classify each subject, i.e., each discourse element under the description, to one or more of the given terms, based on their understanding of what the terms mean.

In the figure, the result from two of the raters is shown. Thus, given e, both rater 1 and rater 2 have included the set {(Board of Directors), (CIO)} in the extension of *"agent"*. Likewise they both agree that the extension of term *"motivates"* is {(Inflation will continue to rise, Expedite hardware renewal program)}. However, we also observe that they do not completely agree on the extension of term "believes", or, reversely, to which term's extension the corresponding elements should belong to. Further, there is an element that has not been part of any extension (the triplet {(Board of Directors, CIO, Expedite hardware renewal program)}), and a term that has empty extensions for both raters (*"plays-role"*). Note that if the description e were different, the classifications of the same domain elements could be different; hence the classifications target subjects (pairs of elements and descriptions) rather than just elements.

By observing the classifications and counting the different kinds of agreement and disagreement incidences as well as the level to which different terms/signifiers and elements participate in ratings we are able to calculate statistics that reveal what kinds of issues the candidate vocabulary has. Following [14], let:

- $I_p(v, e) \subset \mathbf{D}$ be the extension that rater p gave to signifier v under description e.
- $X_p(v) = \bigcup_{e \in E}(I_p(v, e), e)$ be all the subjects that rater p classified under v.

- $B = \{s \in S \,|\, \exists p \in P, \exists v \in V, s.t.\ s \in X_p(v)\}$ be all subjects such that there is a rater that classified them to a signifier.
- $N \subseteq P \times S \times V$ be the set of all instances of a rating of a subject to a signifier by a rater.
- $N_d = \{r \in N |\, d \in \mathbf{D},\ r \text{ includes } d\}$ be the set of all rating instances that rate element $d \in \mathbf{D}$ under some description, to some signifier, by a participant.
- $R(s, v) = \{p \in P | s \in X_p(v)\}$, be the subset of raters that classified subject s under v.

For example, in Figure 1: $I_{p_2}(\text{“}agent\text{”}, e)$ = {(Board of Directors), (CIO)}, $X_{p_1}(\text{“}believes\text{”})$ = {((BD, Inflation will continue to rise), e)}, $R((\text{CIO}, e), \text{“}agent\text{”})$ = $\{p_1, p_2\}$, $R((\text{CIO}, e), \text{“goal”}) = \{\}$, $R((\text{BD, Expedite hardware renewal program}), \text{e}), \text{“}believes\text{”}) = \{p_2\}$, and B contains all subjects except ((BD, CIO, Expedite hardware renewal program), e). Further, N is the number of all arrows in the figure (hence, $|N| = 14$), and $N_{\text{“BD”}}$ are all the arrows in the figure that target element $d =$ “Board of Directors (BD)”, hence $|N_{\text{“BD”}}| = 2$.

Given the above, we can go ahead and define the two types of metrics for measuring within-rater misalignment and *rater-authoritative misalignment*.

3.3 Within-rater misalignment

We first describe the metrics for within-rater misalignment. All these metrics have two dimensions along which they describe the issue. The *intensity* aspect refers to the depth of the issue, i.e. how much a single observation, focusing, e.g., on one subject or element, is indicative of the issue. The *prevalence* aspect refers to how common the issue is in the data, i.e. how common observations that pass the intensity threshold are, e.g., across subjects. Thus, the metrics firstly assume a minimum intensity threshold, at or above which observations are considered relevant evidence of an issue. Subsequently, they measure whether the frequency of such observations exceeds a minimum prevalence threshold.

Construct Deficit. Recall that designers include elements in D if they believe that they are instances of a concept that should be part of the language. It follows that if we observe that a number of elements remain unrated or underrated, the concepts that are expected to classify those elements likely do not have appropriate terms in V that represent them. Hence we have construct deficit.

More formally, let $D_{\texttt{NA}} = \{d \in \mathbf{D} \mid |N_d| \leq t,\ t = 0 \text{ or small}\}$ be the set of elements in $\mathbf{D}$ that receive a number of ratings that is smaller or equal to an intensity threshold t. We have evidence of construct deficit when the size of this set exceeds a prevalence threshold l, i.e., $|D_{\texttt{NA}}| \geq l$, for a small l. By lowering t, the metric becomes less sensitive, as it considers only clear high-intensity cases (e.g., for $t = 0$, only d's with no ratings whatsoever are included in the set). By lowering l the metric instead becomes more sensitive, as it signals deficit even if a smaller set of high-intensity examples (i.e., a smaller number of d's that are underrated) is found.

In our example of Figure 3, for d_1 = (BD, CIO, Expedite hardware renewal program), $|N_{d_1}| = 0$ while for d_2 = (BD, Expedite hardware renewal program), $|N_{d_2}| = 1$. If our intensity and prevalence thresholds are $t = 0$ (the least sensitive) and $l = 0$ (the most sensitive), the metric signals deficit on the basis of d_1. The designers may hypothesize that a term representing a concept such as *delegate* or *order* appears to be missing from the vocabulary. If we set $t = 1$ (more sensitive) and $l = 3$ (less sensitive), although both d_1 and d_2 are now deemed underrated, they are not many enough to signal the problem.

Construct Excess. Recall first that, by its construction, $\mathbf{D}$ must contain a sufficient number of representative instances of each concept that should be included in the language. Hence, each concept should have a non-empty extension over $\mathbf{D}$ for one or more of the descriptions. It follows that if a term is found to be associated with an empty extension, we can infer that it represents none of the concepts of interest and it is hence excessive.

To formalize this, let $v \in V$ a signifier which we wish to investigate if it is excessive. Let $B_v^{\emptyset} = \{s \in B | \, |R(s, v)| \leq t, \text{small } t\}$ be the set of subjects that were generally classified to terms, but the number of raters that classify them under term v specifically is below an intensity threshold t. When almost all subjects are like this, i.e., $|B \setminus B_v^{\emptyset}| \leq l$ for some small prevalence threshold $l \geq 0$, i.e., no or very few of the rated subjects were rated under v, this is evidence of construct excess, with v

being the excessive signifier. As above, by lowering threshold t we decrease the sensitivity of the metric, as we restrict it to the clear high-intensity cases. By lowering l the metric also becomes less sensitive, requiring higher prevalence (i.e., an even smaller proportion of subjects classified under v) before it signals excess.

In our example, there is no subject that has been classified under term v = "plays role", i.e., $\forall s \in B, R(s, \text{“}\textit{plays role}\text{”}) = \{\}$. Hence, $B_v^{\emptyset} = B$, i.e., $|B \setminus B_v^{\emptyset}| = 0$. This is a symptom of *"plays role"* being a superfluous term: i.e., a term that does not represent any concept that is of relevance to this language assuming a representative **D**. In addition, for $t = 1$ (so: slightly higher intensity threshold, which increases sensitivity), the term *"intention"* would be deemed excessive, too: the term is used by only one (intensity) rater (p_1) to classify only one (prevalence) subject: (Expedite hardware renewal program).

Overlap. Before we discuss construct redundancy we first need to describe the notion of a *conceptual overlap* [14]. Assume that for several pairs of raters p_i, p_j, distinct or otherwise, both $s \in X_{p_i}(v_1)$ and $s \in X_{p_j}(v_2)$ for a given subject $s \in S$, i.e., the subject is consistently classified under two different terms by different or the same rater. We then say that there is a conceptual overlap between v_1 and v_2 subject to s.

In our example, considering subject (Expedite hardware renewal program,e), we observe that terms *"goal"* and *"intention"* overlap two times: between the two raters and within rater 1 who classifies the subject under both terms.

Construct Redundancy. The definition of the redundancy construct is based on the observation that terms for which participants tend to form the same extension can be assumed to represent the same concept. One of the terms is hence redundant.

To formalize this, let first $o(s, v_1, v_2)$ be a concrete metric for measuring overlap intensity between v_1 and v_2 with respect to subject s that operationalizes the above principle of conceptual overlap – we will offer a concrete proposal in our study below. The higher $o(s, v_1, v_2)$ the higher the overlap. Let subject $s \in S$ be *relevant* to a signifier v if a minimum number m of raters have included s in their extension of v, i.e., $|R(s, v)| \geq m$. Let L_v be the relevant subjects for v. Further, let $L_{vv'} = L_v \cup L_{v'}$ be the subjects that are relevant to either v or another given concept $v' \neq v$.

Subsequently, define $L^o_{vv'} = \{s \in L_{vv'} \mid o(s, v, v') \geq t\}$ to be the set of subjects relevant to either term v or v' that also exhibit an overlap with respect to those terms that exceeds a given intensity threshold t. If $\exists v' \neq v$, s.t. $|L_{vv'} \setminus L^o_{vv'}| \leq l$, for a small l, i.e., there exists a term v' such that the set of relevant subjects in which v and v' do not overlap is below a prevalence threshold t, this is an indication of construct redundancy for either v or v'.

By lowering the intensity threshold t, we increase the metric's sensitivity, as we consider lower intensity cases (cases of lower overlap) as signals of redundancy. By lowering prevalence threshold l, the metric becomes less sensitive, requiring more evidence of prevalence (i.e., an even smaller proportion of subjects for which v does not overlap with other terms) before it signals redundancy.

In Figure 3, for v = *"intention"*, and assuming relevance threshold $m = 1$, s = ("Expedite h/w renewal program", e) is the only relevant subject, hence, $L_v = \{s\}$. The only v' such that $L_{vv'} \neq \{\}$ is "*goal*" and $L_v = L'_v = L_{vv'}$. Assume, further, that $o(s, v, v') = 0.66$. For $t \leq 0.5$, it follows that $|L_{vv'} \setminus L^o_{vv'}| = 0$, which means that, for $l = 0$ already, v is redundant. The same result emerges from the point of view of v'.

Given that changing the intensity and prevalence thresholds has a different effect to the sensitivity of each metric, it is useful to summarize these effects in Table 1.

	lower t	lower l
Deficit	−	+
Excess	−	−
Redundancy	+	−

Table 1 The effect of *lowering* t and l to metric sensitivity. Compared to a less sensitive metric, a more sensitive metric requires less evidence before signaling the issue.

The above within-rater misalignment detection metrics are abstract in that they are meant to express the principles for measuring each misalignment class in accordance to the definition of the latter. Applications are expected to produce concrete operationalizations of such metrics, depending on the data collection approach and the needs of the study. For example, instruments may

or may not allow raters to apply multiple ratings to a subject. Further, they may ask raters to explicitly mark lack of options (*"None of the above"*) or may infer this decision from their silence. In addition, different studies may have different needs in terms of how they want to present and visualize the metrics and what kind of statistical inference they require. Thus, the metrics allow for flexibility in how they are translated into precise calculations. We propose an example set of such operationalizations in Section 4.3.

3.4 Rater-authoritative misalignment

In addition to constructs that interpret the Wand and Weber misalignment characterization framework, a different set of metrics is also defined for measuring the alignment between users and designers of the language. For the purpose, we assume that the ratings of one of the raters p_a is the authoritative one, i.e., the one that the designers consider to perfectly align with the assumed ontological commitment. Rater-authoritative misalignment is then measured based on the distance between user and authoritative ratings.

Specifically between the authoritative ratings of rater p_a and the ratings of an arbitrary rater p_i we may have with regards to a term v:

Perfect Alignment when $X_{p_i}(v) = X_{p_a}(v)$.

Term Fineness when $X_{p_i}(v) \subset X_{p_a}(v)$. In other words, there is one or more subjects which p_i does not think should be classified under v, but designer p_a thinks they should. As such, signifier v is understood to evoke a more specialized concept than the one the designers intended it to.

Term Coarseness when $X_{p_i}(v) \supset X_{p_a}(v)$. Thus, there is one or more subjects which p_i thinks should be classified under v, but the designers p_a think it should not. As such, the signifier v is understood to evoke a more general concept than the one the designers intended it to.

Unspecified Misalignment when both $X_{p_i}(v) \setminus X_{p_a}(v) \neq \{\}$ and $X_{p_a}(v) \setminus X_{p_i}(v) \neq \{\}$; so neither set of ratings is a subset of the other.

Total Misalignment when $X_{p_i}(v) \cap X_{p_a}(v) = \{\}$, i.e. the two rating sets are disjoint.

When inspecting any case of rater-authoritative misalignment we can measure how broad or narrow the perceived meaning of the signifier is in comparison to its intended meaning. Specifically we have:

Scope Deficiency defined as the difference $X_{p_a}(v) \setminus X_{p_i}(v)$. It includes examples of domain phenomena that a revised (broader) term v' should describe in addition to the ones it currently describes.

Scope Excess (not to be confused with construct excess) defined as the difference $X_{p_i}(v) \setminus X_{p_a}(v)$. It offers examples of domain phenomena that the current term inadvertently describes and which a revised (narrower) term v' should not be perceived to be describing.

Returning to our example, assume that the first rater is the authoritative one $p_a = p_1$. Then, compared to rater p_2, there is perfect alignment for terms such as *"agent"*, *"assessment"*, and *"goal"*. However, there is possible term fineness in *"intention"* with ((Expedite hardware renewal program), e) being the term's scope deficiency: designers think this is an intention but the rater thinks it is not. The opposite, term coarseness, seems to be exhibited with *"believes"* where ((BD, Expedite hardware renewal program), e) is the one subject in the term's scope excess. Designers think that *"believes"* should not evoke a meaning that would allow one to classify the specific subject under the term. However, rater p_2 has done so, assuming probably that the term is adequate for describing the association of an agent (BD) with a goal (Expedite hardware renewal program) – perhaps due to the absence of a more fitting term (e.g., *"wants"*).

Note, finally, that the above metrics compare the authoritative ratings with the ratings of one arbitrary rater, for the purpose of establishing the metrics' meaning. It is up to concrete operationalizations of the metrics to express the exact quantitative formula for multiple raters.

4 Experimental Design

4.1 Overview, Research Questions and Methodology

We have so far presented a set of metrics that we claim can measure, after being properly operationalized, language vocabulary design issues. To empirically test our claims we conducted an experimental study with human participants. The

research questions that we wish to answer with our experiment are as follows:

RQ1: Can the abstractly defined metrics be appropriately operationalized into a reliable data collection instrument?
RQ2: Do the operationalizations successfully detect all known issues with the languages under comparison?
RQ3: Do the operationalizations indicate issues that are known not to exist?
RQ4: Do self-reported measures of deficit, excess, and overlap/redundancy correlate with observational rating-based measures? Further, do self-reported measures independently confirm the expected issues? Finally, is acquisition of self-reporting data reliable?

To answer the above, our experimental design is based on the application of the proposed metrics in two languages. Both languages claim to have vocabularies that are appropriate for modeling human intentions, in the context of, e.g., requirements engineering. However, while the first language is indeed based on an existing language for modeling actor intentions, the second language is an altered version of the first, where specific alignment issues have been purposefully constructed. We then apply the framework on the two languages and address our research questions as follows.

RQ1 is assessed qualitatively on the basis of whether producing an operationalization and data collection instrument was at all possible and indirectly based on whether it led to successful outcomes; i.e., useful observations about the languages. In addition, we perform a reliability test by asking some participants retake the test after a period, and measure the similarity of their two responses. To answer the next two questions, we observe whether application of the metrics will reveal the deliberate/manufactured alignment issues in the second language (**RQ2**), while not offering false indications of issues that were not manufactured or known to exist (**RQ3**). Finally, **RQ4** is assessed through comparison of observational data with attitudinal data and analysis of the latter with respect to the two languages as well as through, again, test-retest reliability analysis.

We continue by presenting the experimental design and related artifacts, the proposed metric implementation, and, finally, the hypotheses to be tested.

4.2 Experimental Artifacts

The experimental artifacts we adopt or develop for this experiment are the two (2) languages under comparison, two (2) world descriptions, and a set of domain elements to be classified under each description.

4.2.1 Languages

The two languages consist of three (3) entities (unary concepts) and four (4) relationships each. The first language, *goal models*, has entity concepts *actor*, *goal*, and *belief* and relationships *wants-to*, *believes-that*, *motivates* and *is-a-means-to*. The language is a subset of iStar 2.0 [9] extended with concepts *belief*, *believes-that*, and *motivates*.

The second language, called *intention models* is the result of changing the first language in a way that some obvious issues are introduced. Firstly, the *actor* concept is replaced with the concept *organization*. The latter is a specialization of the former, and, as such, implies a more restricted extension that excludes, e.g., individual roles and persons. The *belief* concept is also replaced with the quite unrelated *objective* concept. The latter concept has a meaning very similar to that of *goal*, which remains in the second language. The relationship concepts change as follows: *believes-that* is replaced by *intends-to*, which has a similar meaning as *wants-to*. Further, *motivates* is replaced by *prevents*. The corresponding English words are used as terms to describe the concepts – e.g. *"goal"*, *"belief"* etc.

While serving the purpose of this introductory study, the vocabularies can be argued to be relatively small in numbers of concepts, which adds a caveat to generalizing the results to larger languages – we discuss this in Section 6 where we analyze validity threats.

4.2.2 Descriptions and Domain Elements

For the experiment, two world descriptions are presented to experimental participants. The first, e_1, is about Heather, an organic products store owner and her various business concerns, goals,

and decisions. The second, e_2, is about Kim, a sales representative and his various concerns in arranging for his business trips. Both cases are created by the authors. The latter, Kim's case, is inspired in part by an example found in the iStar 2.0 guide [9].

A set of distinguished elements that require to be modeled are identified in each description. The elements identified are ones suited for application of the goal modeling language, according to the authors' opinion. Thus, we expect that the first language will not exhibit any of the issues introduced earlier. Of the elements, 24 (twenty-four, 12 in each case) correspond to entities and 21 (twenty-one, 9 from Heather's case and 12 from Kim's case) correspond to binary relationships (i.e., they are tuples). The elements are distinct across cases, so the set of all subjects maps 1-1 to **D**. The authors also define the *authoritative* way by which the elements are supposed to be classified/modeled. Their distribution can be seen in Table 2. The authoritative responses are the same between the two languages. For example, the authoritative responses for *"organization"* in intention models are the same as those of the term from goal models it replaces (*"actor"*). In other words, we assume that the designers of the language for intention models believe that *"organization"* is a good term for representing concept *actor*.

Heather's Case		Kim's Case		
Concept	#	Concept	#	Total
Actor	1	Actor	4	5
Goal	6	Goal	5	11
Belief	5	Belief	3	8
wants-to	3	wants-to	4	7
believes-that	2	believes-that	2	4
motivates	2	motivates	3	5
is-a-means-to	2	is-a-means-to	3	5

Table 2 Number of elements (#) by case and authoritative designation.

4.2.3 Instruments and Process

Given the languages, the two world descriptions and the elements, the experimental instruments for acquiring participant ratings and attitudinal data are developed. The instruments are survey-like on-line sequences of screens, and in each screen the participants are presented with information and/or asked for their input. Psytoolkit is used to develop and host the instruments [20, 21]. Two such instruments are developed, one for each language. Participants are assigned to each language/instrument randomly and in a *between-subjects* manner. The following are the main contents of the instruments in this sequence:

Video Presentation: Participants first watch a video presentation of the corresponding language (4:46 minutes for goal models, 4:40 minutes for intention models). Concepts are presented with a definition and examples. We discuss the choice of training time and administration method from the viewpoint of internal and external validity in Section 6.

Comprehension Test: Subsequently, participants respond to a set of questions assessing attendance and comprehension of the videos. They are asked to match definitions and examples offered in the videos of the preceding screen with the corresponding concepts. More than three (3) erroneous responses out of the six (6) plus eight (8) questions about entities and relationships for which an authoritative correct exists, respectively, disqualifies the participant.

Main Rating Exercises: In two subsequent screens, Heather's and Kim's cases are presented on top of each page. Below the description, the elements that require rating are listed. For simplicity, the ones that require an entity rating are separated from the ones that require a relationship rating; so the authors designate what element should be classified under a unary term and what under a binary one. Below each element the inventory of entities or relationships of the language in question is offered, augmented with the "`None of the above`" (henceforth "None" for brevity) option and using square select-all-that-apply type check-box widgets. For example ''*Heather*" is presented as a concept element below which the inventory {Actor, Goal, Belief, None} is added in the goal model instrument and the inventory {Organization, Goal, Objective, None} in the intention model one. Thus, other than the inventories, for each description (Heather's case vs. Kim's case), the two rating pages (the one presented to the goal models group and the one shown to the intention models group) are identical in terms of descriptions and elements to be classified.

However, the elements in each category (unary concepts and binary relationships) are provided in a random order.

Self Reporting: After the video presentations and both before and after the rating exercises are completed, participants are asked to evaluate the language in various ways. In three (3) separate screens participants are asked to: (a) rate from 0 to 10 (default 5) the *"relevance"* of each concept in the language they worked with, (b) rate from 0 to 10 completeness of the entities and, separately, the relationships of the language (i.e., whether they thought that *"the set of concepts included in the language [...] were sufficient for characterizing relevant parts in the described cases"* and that *"No more [concepts] need to be added to the set to make it more complete"*, (c) for each pair of entities and each pair of relationships, rate from 0 to 10 the *conceptual overlap* between the members of the pair, i.e., the extent to which they *"refer to the same thing"*. We note that the overlap questions screen (c) precedes the classification exercises, whereas the other two questions succeed the exercises in the task ordering.

Demographics: In the last screen participants offer demographic information.

Participation is solicited from the on-line participant pool Prolific [22, 23]. Participants are required to be based in an English speaking country (UK, US, Canada, Australia, New Zealand) and have an undergraduate degree in Computer Science or Computing (IT), according to self-reported information. The input from a total of 30 participants is solicited, 15 for each condition/language. The rationale for the targeted sample size is purely pragmatic: we want to investigate if a language designer with a relatively modest budget for participant inducements (our main cost, excluding re-test fees, was about £200, in 2022 prices) has enough statistical power to identify at least some statistically significant results.

4.3 Metric Implementations

Let us now discuss metric implementations that are suitable for the particular instruments and the data they offer. Let us first observe that the two languages, goal models (gm) and intention models (im) have vocabularies (arities in parentheses) V_{gm} = {*"actor"(1)*, *"goal"(1)*, *"belief"(1)*, *"wants-to"*(2), *"believes-that"*(2), *"motivates"*(2), *"is-a-means-to"*(2)} and V_{im} = {*"organization(1)"*, *"goal(1)"*, *"objective(1)"*, *"wants-to"*(2), *"intends-to"*(2), *"prevents"*(2), *"is-a-means-to"*(2)}. There is a set of two descriptions $E = \{e_1, e_2\}$, corresponding to Heather's and Kim's cases, and a total of 45 domain elements and tuples thereof are defined $\mathbf{D}$ = {(Kim), ((Heather), (introduce a loyalty program)),...}. Given that each element of $\mathbf{D}$ appears in only one description, a total of 45 subjects are rated by two groups P_{gm} and P_{im} of participant raters assigned to each language.

Let function $n : P \times S \times V \mapsto \{0, 1\}$, be $n(p, s, v) = 1$ if rater $p \in P$ has classified $s = (d, e)$ under $v \in V$, and $n(p, s, v) = 0$ otherwise. Consider also an additional special term $v_{\mathtt{NA}}$ representing the "None" rating.

Denote the marginal sums:

$$n(\cdot, s, v) = \sum_{p \in P} n(p, s, v)$$

$$n(p, \cdot, v) = \sum_{s \in (\mathbf{D} \times E)} n(p, s, v)$$

$$n(p, s, \cdot) = \sum_{v \in V} n(p, s, v)$$

Such sums over more than one variable, or over constituents of subjects (elements or descriptions), are understood normally. For example $n(p, (d, \cdot), \cdot) = \sum_{e \in E}(\sum_{v \in V} n(p, (d, e), v))$, representing the set of ratings in which participant p classifies element d under some term and $n(\cdot, (d, \cdot), \cdot) = \sum_{p \in P} n(p, (d, \cdot), \cdot)$ is the set of all ratings on subjects that mention element d. Notice that the above summations over V do not include $v_{\mathtt{NA}}$ as $v_{\mathtt{NA}} \notin V$.

Then we can implement the metrics we discussed earlier as described in the following subsections. All implementations are based on a common idea. Firstly, a numerical value for representing per-observation intensity is defined, allowing for the construction of a set of intensity values for all observations. Then, quantiles are used for the measurement of prevalence allowing us to identify the lowest or highest occurred intensity of the issue after the exclusion of outliers.

4.3.1 Construct Deficit

Recall, first, that, by metric definition, we have evidence of construct deficit when $|D_{\texttt{NA}}| \geq l$, for a small threshold l, where $D_{\texttt{NA}}$ is the set of elements whose number of ratings is below a threshold, $D_{\texttt{NA}} = \{d \in \mathbf{D} \mid |N_d| \leq t,\ t = 0 \text{ or small}\}$. In the instrument used in this experiment, participants explicitly rate subjects as unclassifiable through ratings to the special term $v_{\texttt{NA}}$. Hence, let $N_d^{\texttt{NA}}$ be the set of classification instances in which d is classified in $v_{\texttt{NA}}$. For clarity, N_d represents non-NA classifications of d, and the two sets N_d and $N_d^{\texttt{NA}}$ are, hence, disjoint.

To construct our operationalization we need a measure of $|N_d|$ being small and, subsequently, a measure of $|D_{\texttt{NA}}|$ being large. For the former, rather than using $|N_d|$ in absolute terms, we compare $|N_d|$ to $|N_d| + |N_d^{\texttt{NA}}|$ by forming the ratio of the two. The result, which lies in the interval [0,1], is subtracted from 1 so that the greater the value the more the evidence of deficit. Hence: $1 - |N_d|/(|N_d| + |N_d^{\texttt{NA}}|) = |N_d^{\texttt{NA}}|/(|N_d| + |N_d^{\texttt{NA}}|)$. Given the form of our instrument, $|N_d| = n(\cdot, (d, \cdot), \cdot)$ and $|N_d^{\texttt{NA}}| = n(\cdot, (d, \cdot), v_{\texttt{NA}})$.

The above handles intensity. To measure prevalence we consider the set of intensity values for all $d \in \mathbf{D}$ and examine the maximum, if one deficit-intense d suffices as evidence (i.e., $l = 1$), or a large quantile, if we demand more than one such d's to exist before we conclude deficit of the vocabulary. Hence the final metric is:

$$\text{def}_V = Q^c(\{\frac{n(\cdot, (d, \cdot), v_{\texttt{NA}})}{n(\cdot, (d, \cdot), \cdot) + n(\cdot, (d, \cdot), v_{\texttt{NA}})} \mid d \in \mathbf{D}\})$$

where $Q^c(X)$ is the c-th percentile of the set X, e.g., $c = 100$ (so: max) or $c = 90$ (90th percentile). We also define the per participant deficit using the same ratio:

$$\text{def}_V^{pp}(p) = Q^c(\{\frac{n(p, (d, \cdot), v_{\texttt{NA}})}{n(p, (d, \cdot), \cdot) + n(p, (d, \cdot), v_{\texttt{NA}})} \mid d \in \mathbf{D}\})$$

Per-participant deficit allows for statistical analyses. In our case, we compare two independent samples, $\mathbf{def}_{V_{gm}}^{pp} = \{\text{def}_{V_{gm}}^{pp}(p) \mid p \in P_{gm}\}$ and $\mathbf{def}_{V_{im}}^{pp} = \{\text{def}_{V_{im}}^{pp}(p) \mid p \in P_{im}\}$ of, we assume, independent and identically distributed (i.i.d.) values.

4.3.2 Construct Excess

Recall that we defined B to be the set of all subjects that have been classified to some signifier and $B_v^{\emptyset}$ the subset of B containing subjects that no or few participants classified specifically under v: $B_v^{\emptyset} = \{s \in B \mid |R(s, v)| \leq t, \text{small}\, t\}$. We, then, observed that if it is actually all or most of the subjects in B that enjoy few or no classifications under v, i.e., $|B \setminus B_v^{\emptyset}| \leq l$ for some small $l \geq 0$, this is evidence of v's construct excess: v is not used for classifications.

As above, let us first construct the intensity aspect. Observe that in our case $|R(s, v)| = n(\cdot, s, v)$ (number of raters that classified s under v) and is bounded by $|P|$ (total number or raters). Set then $t' = t/|P|$. If the quantity $U(s, v) = n(\cdot, s, v)/|P| \leq t'$ for some s, this is evidence of v being excessive with respect to s. Hence, $U(s, v)$ is our intensity metric (the lower it is, the more the excess intensity with respect to s, v).

Prevalence is then measured by the extent to which the number of subjects in S for which $U(s, v)$ is small (i.e., $B_v^{\emptyset}$) approaches the total number of subjects rated (i.e., B). We again use a percentile to see if this is observed in few, most or all subjects. Hence, the comprehensive metric:

$$\text{exc}_V(v) = 1 - Q^c(\{U(s, v) \mid s \in S\})$$

... is close to 1 when v is excessive. Q^c is the c-th percentile as above – likely $c = 100$ in practice, or lower for less tolerance with respect to prevalence.

A similar construct can also be defined on a per-participant basis. Recall that $n(p, \cdot, v)$ is the total number of subjects that participant p rated under term v – bounded now by the total number of subjects $|S|$. Then:

$$\text{exc}_V^{pp}(p, v) = 1 - \frac{n(p, \cdot, v)}{|S|}$$

Finally, for a per-participant measure of the excess of an entire vocabulary $v \in V$ we can simply calculate some statistic over the set of excess values of individual constituent concepts. Using the mean as an example:

$$\text{exc}_V^{pp}(p) = mean(\{exc_V^{pp}(p, v) \mid v \in V\})$$

As above, in our example, samples $\mathbf{exc}^{pp}_{V_{gm}} = \{\text{exc}^{pp}_{V_{gm}}(p) \mid p \in P\}$ and $\mathbf{exc}^{pp}_{V_{im}} = \{\text{exc}^{pp}_{V_{im}}(p) \mid p \in P\}$ can be used for statistically comparing the construct excess of two vocabularies in their entirety, while samples $\mathbf{exc}^{pp}_{V_{gm}}(v) = \{\text{exc}^{pp}_{V_{gm}}(p, v) \mid p \in P\}$ and $\mathbf{exc}^{pp}_{V_{im}}(v') = \{\text{exc}^{pp}_{V_{im}}(p, v') \mid p \in P\}$, allow comparison of term v with, e.g., a candidate replacement v'.

4.3.3 Construct Redundancy

Let us first calculate overlap between signifiers v_1 and v_2 on the basis of pairwise disagreements involving the two concepts over the maximum such disagreements can possibly be. The following can be shown to be an adequate measure of that:

$$o_V(s, v_1, v_2) = \frac{n(\cdot, s, v_1) \times n(\cdot, s, v_2)}{\lfloor n(\cdot, s, \cdot)/2 \rfloor \times \lceil n(\cdot, s, \cdot)/2 \rceil}$$

Let $L_{vv'} = L_v \cup L_{v'} = \{s \in S | ((n(\cdot, s, v) \geq \alpha \times n(\cdot, s, \cdot)) \vee (n(\cdot, s, v') \geq \alpha \times n(\cdot, s, \cdot)))\}, \alpha \in [0, 1]$, be the set of subjects in S for which at least a fraction α of the total classifications they received was under v or v'; i.e. the set of subjects that are *relevant* with respect to either v or v'.

Let then $\mathbf{ov}_V(v, v') = \{o_V(s, v, v') \mid s \in L_{vv'}\}$ be the set of overlap measures between v and v' over all relevant subjects. Recall that redundancy is measured by the prevalence of high overlaps, specifically the extent to which there are elements of $\mathbf{ov}_V(v, v')$ that are small, for some $v' \neq v$. Hence, as above, construct redundancy for v can be measured by:

$$\mathbf{rdn}_V(v) = \max_{v' \in \mathbf{V} \setminus \{v\}} \{Q^c[\mathbf{ov}_V(v, v')]\}$$

i.e., the maximum overlap exhibited in comparison to every other construct, measured as the minimum (c = 0) or other low percentile of the elementary overlaps that occurred between v and the other construct. In other words, we develop a set of overlap values of v with every other signifier v'; the values represent the intensity aspect. Then, through the quantile, we assess prevalence by examining if most or all of those values are actually high.

Note that the mean (or other statistic) over the elements of $\mathbf{ov}_V(v, v')$ offers a quick descriptive indicator of overlap $\text{ov}_V(v, v')$.

As above, we can define overlap per participant, if the instrument allows multiple ratings per subject, as is our case here. Firstly the elementary overlap in the ratings of a participant on a subject is again defined as:

$$o_V(p, s, v_1, v_2) = \frac{n(p, s, v_1) \times n(p, s, v_2)}{\lfloor n(p, s, \cdot)/2 \rfloor \times \lceil n(p, s, \cdot)/2 \rceil}$$

and likewise the overlap between v and v' according to participant p can be the average or other statistic of the observed overlaps over subjects, such as:

$$\text{o}^{pp}_V(p, v, v') = mean(\{o(p, s, v, v') \mid s \in S\})$$

Note that here we consider all subjects, not just the relevant ones. Given this construct we can compare, e.g., term v' in language *gm* versus its alternative v'' in language *im* with respect to its overlap to v by comparing the samples $\mathbf{ov}^{pp}_{V_{gm}}(v, v') = \{\text{o}^{pp}_{V_{gm}}(p, v, v') | p \in P_{gm}\}$ with $\mathbf{ov}^{pp}_{V_{im}}(v, v'') = \{\text{o}^{pp}_{V_{im}}(p, v, v'') | p \in P_{im}\}$, assuming that the two competing languages are rated by distinct groups of raters P_{gm} and P_{im} to be able to assume independence.

4.3.4 Accuracy

Given the set of authoritative ratings, we can also define three functions:

$$acc(p, s, v) = \begin{cases} \mathbf{1}, \text{ if } n(p_a, s, v) = 1 \text{ and } n(p, s, v) = 1 \\ \mathbf{0}, \text{ otherwise} \end{cases}$$

$$def(p, s, v) = \begin{cases} \mathbf{1}, \text{ if } n(p_a, s, v) = 1 \text{ and } n(p, s, v) = 0 \\ \mathbf{0}, \text{ otherwise} \end{cases}$$

$$exc(p, s, v) = \begin{cases} \mathbf{1}, \text{ if } n(p_a, s, v) = 0 \text{ and } n(p, s, v) = 1 \\ \mathbf{0}, \text{ otherwise} \end{cases}$$

The marginal totals as per the above notation $acc(\cdot, \cdot, v)$, $def(\cdot, \cdot, v)$, and $exc(\cdot, \cdot, v)$ offer a raw measure of the *accuracy*, *deficiency* and *scope excess* of a given term. For example $exc(\cdot, \cdot, v) = \sum_{p \in P} \sum_{s \in S} exc(p, s, v)$ measures the total number of ratings that involved concept v, when the authoritative rating was $v' \neq v$. Further, $acc(p, \cdot, v)$, $def(p, \cdot, v)$, and $exc(p, \cdot, v)$ count the corresponding raw occurrences for a single participant p while for the whole vocabulary we have the marginals $acc(\cdot, \cdot, \cdot)$, $def(\cdot, \cdot, \cdot)$, and $exc(\cdot, \cdot, \cdot)$. For descriptive analysis, the raw numbers can be used

to develop Euler diagrams for visualizing the quality and level of misalignment. We show examples below.

Additional indicators of the accuracy of the language that can be derived from the above are *precision* and *recall*. Precision (prec) measures the proportion of rater classifications that are in agreement with the authoritative, while recall (rec) measures the proportion of authoritative classifications that are also performed by raters. Hence, focusing on term v, while precision is the number of "true positives" $acc(\cdot,\cdot,v)$ over the total number of "positives" $acc(\cdot,\cdot,v) + exc(\cdot,\cdot,v)$, recall is the number of "true positives" over the total number of "true" elements $acc(\cdot,\cdot,v) + def(\cdot,\cdot,v)$:

$$\text{prec}_V(v) = acc(\cdot,\cdot,v)/(acc(\cdot,\cdot,v) + exc(\cdot,\cdot,v))$$
$$\text{rec}_V(v) = acc(\cdot,\cdot,v)/(acc(\cdot,\cdot,v) + def(\cdot,\cdot,v))$$

The per-participant metrics, amenable to statistical comparisons are:

$$\text{prec}^{pp}_V(p,v) = acc(p,\cdot,v)/(acc(p,\cdot,v)+exc(p,\cdot,v))$$
$$\text{rec}^{pp}_V(p,v) = acc(p,\cdot,v)/(acc(p,\cdot,v) + def(p,\cdot,v))$$

And similarly we can define metrics for the entire vocabulary on a per-participant basis:

$$\text{prec}^{pp}_V(p) = acc(p,\cdot,\cdot)/(acc(p,\cdot,\cdot) + exc(p,\cdot,\cdot))$$
$$\text{rec}^{pp}_V(p) = acc(p,\cdot,\cdot)/(acc(p,\cdot,\cdot) + def(p,\cdot,\cdot))$$

In our example, sets such as:

$$\mathbf{rec}^{pp}_{V_{gm}}(\text{"}goal\text{"}) = \{\text{prec}^{pp}_{V_{gm}}(p, \text{"}goal\text{"}) \mid p \in P\}, \text{ or}$$
$$\mathbf{prec}^{pp}_{V_{im}} = \{\text{prec}^{pp}_{V_{im}}(p) \mid p \in P\}$$

can be used for statistical inferences about the recall or precision of a term or of an entire vocabulary, respectively.

4.4 Hypotheses

Having defined the operationalizations we are now in the position to express our experimental hypotheses. Based on how we engineered the two languages, we devised a number of such hypotheses about the metrics' results that have to be true if the metrics indeed detect the manufactured issues and only those issues. A summary of the hypotheses can be viewed in Tables 6, 9 and 11. Each of the identified hypotheses is evaluated qualitatively and, whenever available, through inferential statistics.

Let us first offer a high-level description of what we expect to observe given the manipulations we performed to the goal modeling language in order to produce the intention modeling one, and how these expectations can be translated into the testable hypotheses.

General indicators. In general, the manufactured intention models are expected to have an overall lower accuracy (**A.1** and **A.2** in Table 6), higher deficit (**D.1.**, **D.2.** in Table 9), due to the absence of the *"belief"* and *"believes-that"* concepts, as well as higher excess **E.1** due to *"prevents"* and possibly due to the introduced overlaps with ''*objective"* and ''*intends-to"*. We are next looking at metrics concerning the specific interventions we made in the language.

Turning *"actor"* to *"organization"*. With this change we use a term with a narrow extension to describe a concept with a broader one. As such, we expect the construct to exhibit higher levels of scope deficiency and, likewise, lower levels of recall than the *"actor"* counterpart (**A.3** in Table 6).

Replacing *"belief"* with *"objective"*. With this change we use a term to describe phenomena that should better be described by the original term. We expect, thus, the new term to result to less precision and recall compared to the term it replaces (**A.4**). In addition, the new term now overlaps with *"goal"*. We, thus, expect such overlap to be observed (**O.1** in Table 11) and be higher than the one between *"goal"* and *"belief"* (**O.2**). The concepts with the overlap are also likely to exhibit high redundancy (**O.3**).

Replacing *"believes-that"* with *"intends-to"*. As above, the latter term will exhibit lower accuracy (**A.5**), high overlap with *"wants-to"* (**O.1 - O.2**) and high redundancy (**O.3**).

Replacing *"motivates"* with *"prevents"*. This is similar to the previous two replacements with the additional issue of excess, as there are not examples of a *prevents* concept in the elements set and descriptions. Specifically we expect the element to demonstrate high scope deficiency and low recall (**A.6**) compared to the term it replaces (i.e., few will guess – rightly – that *"prevents"* is a term that describes concept *motivates*). It will also present high excess (**E.2**, **E.3**).

Keeping *"is-a-means-to"* as is. We expect that all measures related to the concept are similar between the two languages (**A.7**, **E.4**). Note that while *"goal"* and *"wants-to"* also stay the same, metrics relating to them may still be affected by their overlap with *"intention"* and *"intends-to"*.

	Concept	Recall	Precision
V_{gm}	Actor	100 [100,100]	91.55 [82.08,96.05]
	Goal	90.21 [84.74,93.85]	92.81 [87.7,95.86]
	Belief	91.35 [85.76,96.04]	82.61 [76.12,88.15]
V_{im}	Organization	71.43 [60.81,80]	70.42 [60,79.87]
	Goal	40.26 [33.25,46.96]	79.49 [68.63,87.85]
	Objective	22.32 [15.12,31.17]	17.73 [12.26,25.12]

Table 3 Accuracy measures for entities (%) for goal models V_{gm} and intention models V_{im}. In brackets the 2.5th and 97.5th percentiles of the bootstrapping distribution.

	Concept	Recall	Precision
V_{gm}	wants-to	93.41 [88.54,96.95]	93.41 [88.54,96.95]
	believes-that	96.15 [92.67,99.90]	75.76 [65,82.96]
	motivates	84.62 [73.83,91.57]	71.43 [62.33,80.14]
	is-a-means-to	90.77 [80.36,95.88]	80.82 [72.28,87.28]
V_{im}	wants-to	77.55 [68.79,85.09]	73.79 [65.64,81.18]
	intends-to	8.93 [3.16,19]	5.43 [2.02,11.13]
	prevents	8.57 [3.9,14.93]	46.15 [17.84,73.22]
	is-a-means-to	92.86 [85.61,97.38]	69.89 [61.69,77.99]

Table 4 Accuracy measures for relationships (%). In brackets the 2.5th and 97.5th percentiles of the bootstrapping distribution.

Self reported data. When comparing observational with self-reported data we should observe that there is a negative correlation between deficit and reported completeness (**S.1**), a negative correlation between excess and reported relevance for each term (**S.2**) and a positive correlation between the observed and reported overlap (**S.3**). In addition we test if the self-reported data independently reveal the expected differences between the languages with regards to deficit, excess and overlaps (**S.4**).

Test-retest reliability. We finally perform a test-retest reliability analysis focused on the goal models, whereby the participants are called back to redo the same test, and the agreement between the two responses is measured. We expect that the level of agreement is substantial for the majority of participants, supporting the reliability of the instrument and process (**T.1**).

Of all the hypotheses (Tables 6, 9 and 11), those in the groups **A**, **D**, **E** and **O** attempt to answer **RQ2**, except for hypotheses **A.7** and **E.4** which address **RQ3**, while hypotheses in **S** address **RQ4**. Of all these hypotheses, twenty-three (23) are tested statistically, and given that they are all based on the same data, we perform a Bonferroni adjustment to our Type I error threshold. To allow for equal treatment of all measured aspects the family-wise threshold, 0.05, is firstly shared equally to the five (5) groups of hypotheses where statistical tests are planned (corresponding to letters **A**, **D**, **E**, **O**, **S**). Then in each group, the discounted threshold $\alpha = 0.05/5 = 0.01$ is further divided by the number of hypotheses tested for each group. For example, for accuracy (**A**), we conduct ten (10) experiments with an alpha level of $0.01/10 = 0.001$. Independent sample Wilcoxon rank-sum tests are reported (statistic W, p value, and effect size r) unless otherwise noted. Several hypotheses, however, do not allow for statistical analysis and are hence explored descriptively. Note also that **RQ1**, which concerns the overall feasibility of the endeavor, is responded to qualitatively at the end of the analysis, while the reliability aspect embedded in that research question is addressed through the test-retest analysis (**T.1**).

5 Results[2]

The input from thirty (30) participants is solicited, and fifteen (15) are assigned to each treatment. Of those who took part, two (2) and one (1) participants are excluded in each of the two treatments, goal and intention models, respectively, due to them failing three (3) or more simple video comprehension questions. The gender and age characteristics of participants can be found in Table 5.

Condition	Sex	Average Age	Count
Goal Models	Female	29	4
	Male	27	9
Intention Models	Female	26	2
	Male	26	12

Table 5 Participants Counts and Ages by Condition and Sex.

In the rest of the section we present the results on accuracy, deficit, excess, redundancy, relationship to self reported data, and the test-retest analysis in that order.

5.1 Accuracy

The results for accuracy can be found in Tables 3 and 4 for entities and relationships, respectively. In the tables, the recall and precision measures

[2] All acquired data as well as analysis scripts in R can be found in the accompanying reproducibility package [24].

	H_1	Description	Tests	Outcome
Accuracy	**A.1.**	The vocabulary-wide accuracy metrics will be higher for goal models than for intention models. Pattern exhibited in individual terms except *"is-a-means-to"*.	$\text{prec}_{V_{gm}}(v_1) > \text{prec}_{V_{im}}(v_2)$ and $\text{rec}_{V_{gm}}(v_1) > \text{rec}_{V_{im}}(v_2)$ except *"is-a-means-to"* for $v_1 \in V_{gm}, v_2 \in V_{gm}$ except *"is-a-means-to"*	☆ (with comments)
			$\text{prec}_{V_{gm}} > \text{prec}_{V_{im}}$ and $\text{rec}_{V_{gm}} > \text{rec}_{V_{im}}$	☆
	A.2.	Per-participant accuracy higher for goal models.	$\mathbf{prec}^{pp}_{gm} >_m \mathbf{prec}^{pp}_{im}$	☆★
			$\mathbf{rec}^{pp}_{gm} >_m \mathbf{rec}^{pp}_{im}$	☆★
	A.3.	*"Organization"* has less per-participant recall and more deficiency than "actor".	$\mathbf{rec}^{pp}_{gm}(\text{"}actor\text{"}) >_m \mathbf{rec}^{pp}_{im}(\text{"}organization\text{"})$	☆★
			"organization" has higher deficiency than *"actor"*	☆
	A.4.	*"Belief"* has higher per-participant precision and recall than *"objective"*.	$\mathbf{prec}_{gm}(\text{"}belief\text{"}) >_m \mathbf{prec}_{im}(\text{"}objective\text{"})$	☆★
			$\mathbf{rec}_{gm}(\text{"}belief\text{"}) >_m \mathbf{rec}_{im}(\text{"}objective\text{"})$	☆★
	A.5.	As above between *"believes-that"* and *"intends-to"*.	$\mathbf{prec}^{pp}_{gm}(\text{"}believes\text{-}that\text{"}) >_m \mathbf{prec}^{pp}_{im}(\text{"}intends\text{-}to\text{"})$	☆★
			$\mathbf{rec}^{pp}_{gm}(\text{"}believes\text{-}that\text{"}) >_m \mathbf{rec}^{pp}_{im}(\text{"}intends\text{-}to\text{"})$	☆★
	A.6.	In intention models *"prevents"* exhibits low recall compared to *"motivates"*.	$\mathbf{rec}^{pp}_{gm}(\text{"}motivates\text{"}) >_m \mathbf{rec}^{pp}_{im}(\text{"}prevents\text{"})$	☆★
	A.7.	All accuracy measures of *"is-a-means-to"* remain about the same across languages.	$\mathbf{prec}_{V_{gm}}(v) \simeq \mathbf{prec}_{V_{im}}(v)$ $\mathbf{rec}_{V_{gm}}(v) \simeq \mathbf{rec}_{V_{im}}(v)$	(☆✪)

Table 6 The experimental hypotheses: accuracy
a $>_m$ **b** is interpreted into a one-tail Wilcoxon test for independent samples **a** and **b**
Outcomes: observed qualitatively (☆), and, where statistical test is applicable, passes (★), fails (✪)

for each concept are given. The 95% confidence interval presented in the brackets is created by bootstrapping [25] accuracy, excess and deficiency values 1000 times and conservatively calculating accuracy and precision based on the corresponding quantiles of the collected values.

Recall that we expect that goal model concepts will evoke more accurate responses for most concepts and the overall metric. We find this to be the case for most concepts and collectively the difference to be qualitatively salient (**A.1** – 85.1% and 92% vs. 48.9% and 45.9% precision and recall for goal models and intention models, respectively) and statistically significant (**A.2** – Wilcoxon rank sum $W = 168, p < 0.001, r = 0.72\,(large)$ and $W = 179, p < 0.001, r = 0.82\,(large)$ for precision and recall on the per-participant metrics, respectively). The only possible exception is the high precision value for *"goal"* in intention models: we expected that many elements authoritatively defined as *"goal"* will be in fact classified as *"objective"*, yielding lower precision than the one observed. In fact, it appears that of the two overlapping terms the first one was used the most. Further, *"organization"* appears to show a moderately high accuracy in intention models. We rather expect a relatively low recall and a precision potentially as high as that of *"actor"* in goal models. A look into the data shows that many participants rate subjects such as *"Heather"*, *"Supervisor"* as *"organizations"*, hence, the unexpectedly high recall. The lower precision can be attributed to the facts that there is only one subject in the set that is truly an *"organization"* and precision depends on it alone collecting all accurate responses. This emerges in excess as well – more below.

Pairwise comparisons between concepts also turn out as expected. Firstly, term *"organization"* has lower recall than the concept it replaces (**A.3** – $W = 143, p < 0.001, r = 0.6\,(large)$). Further, the hypothesized (**A.4**) precision and recall differences between *"belief"* and *"objective"* are found to be statistically significant $W = 181, p < 0.001, r = 0.85\,(large)$ and $W = 178.5, p < 0.001, r = 0.84\,(large)$ and so are the differences between *"believes-that"* and its replacement *"intends-to"* for both precision and recall (**A.5** – $W = 181, p < 0.001, r = 0.89\,(large)$ and $W =$

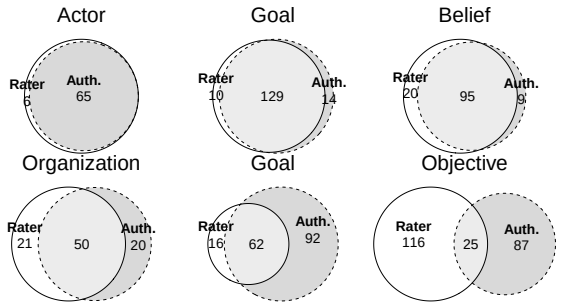

Figure 4 Euler accuracy diagrams for entities

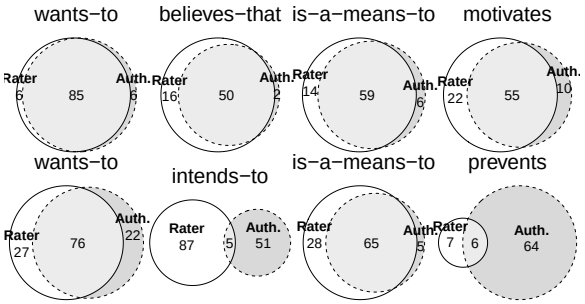

Figure 5 Euler accuracy diagrams for relationships

$182, p < 0.001, r = 0.91\,(large)$). Lastly, the *"prevents"* relationship, which has been deliberately added to trigger construct excess, is found to have, as expected, low recall compared to *"motivates"* (**A.6** – $W = 172, p < 0.001, r = 0.81\,(large)$).

Finally, the relationship *"is-a-means-to"* shows some differences between the two languages which however, do not appear to be substantial compared to the other concepts (**A.7**). A $\alpha = 0.05$ equivalence test for a large effect size (1) fails, which is unsurprising given our small sample size.

A graphical view of the accuracy result can be rendered in the form of Euler accuracy diagrams such as those of Figures 4 and 5, where the scope fineness and scope excess of each term can also be reviewed.

5.2 Construct Deficit

	Type	max	q95	q90	q75
V_{gm}	Entities	0.29	0.14	0.11	0.07
	Relationships	0.07	0.07	0.06	0.00
V_{im}	Entities	0.86	0.71	0.67	0.36
	Relationships	1.00	0.60	0.50	0.35

Table 7 Deficit measures for goal models (V_{gm}) and intention models (V_{im}): maximum observed, and 95th, 90th and 75th quantile.

We hypothesize that intention models will exhibit higher deficit than goal models. This is due to the absence terms *"belief"* and *"believes-that"* in intention models. Recall that, to measure construct deficit, we use the rate by which the "None" option is selected by participants for each element and we identify the maximum and/or upper quantiles across all these rates. The result can be seen in Table 7. The values in the table clearly indicate the increased deficit in intention models (**D.1**), in both intensity (how big the numbers are) and prevalence (how persistent high numbers are as quantiles lower).

We can, further, statistically compare the per-participant deficit between the two languages (**D.2**). Indeed, for both entities ($W = 19, p < 0.001, r = 0.69\,(large)$) and relationships ($W = 0, p < 0.001, r = 0.87\,(large)$) intention models exhibit more deficit, with difference between means 0.206 and 0.174, respectively.

5.3 Construct Excess

The construct excess calculated for each of the concepts of the languages can be viewed in Table 8. Recall that here we expect *"prevents"* to be clearly excessive and, further, due to this concept and the introduced overlaps, the entire intention models language to be more excessive as well. The indicators are as expected, though the statistical test comparing the two languages does not meet our discounted alpha threshold $W = 132, p = 0.0246$ (**E.1**). That there is only one clearly excessive term may be a contributor to this. In terms of individual excess measures, the highest is for *"prevents"* in intention models (0.5) while non-zero indications appear also in concepts *"goal"*, *"objective"*, *"wants-to"* and *"intends-to"* due to the overlaps in which some participants consistently use mostly one of the concepts in the overlapping pair. These are, again, expected (**E.2**).

We can further preform statistical comparisons of the per-participant excess for each of the pairs of corresponding concepts of the languages. The comparison between *"motivates"* and *"prevents"*, which is of interest here (**E.3**) yields a significant result – $W = 0.5, p < 0.001, r = 0.87\,(large)$. However, so is one more comparison (*"goal"*, between the two languages), which can be explained due to its overlaps with *"objective"*. In other words, if an

Goal Models		Intention Models		
Concept	Exc	Concept	Exc	p
Goal	0.00	Goal	0.36	0.12
Belief	0.00	Objective	0.07	0*
Actor	0.00	Organization	0.14	0.23
wants-to	0.00	wants-to	0.14	0.82
believes-that	0.00	intends-to	0.14	0.04
motivates	**0.08**	**prevents**	**0.50**	**0***
is-a-means-to	0.00	is-a-means-to	0.00	0.18

Table 8 Excess per concept. (*) = comparison between line items (e.g. *Actor* vs. *Organization*) significant at $p = 0.0033/7 = 0.00048$ after Bonferroni correction for the 7 comparisons – significance tests over per-participant measures.

overlap is observed, then the concepts participating in the overlap may naturally exhibit increased excess.

We finally observe that the metric does not seem to be picking up false positives (**E.4**) with one exception. All excess measures of goal models are zero as expected and so is intention model concept *"is-a-means-to"*, exactly as expected. However, *"organization"* appears to exhibit some excess in intention models, which is harder to explain than the similar effects in *"goal"* or *"wants-do"* – which are likely due to the overlapping relationships with *"objective"* and *"intends-to"*, respectively. Looking into the data, we find that this is due to the fact that there is only one true *organization* in the elements ("On-line travel agency"), and, hence, only one chance for the signifier to attract enough participant ratings to be deemed non-excessive. This did not happen in our data: there were some "detractors" with respect to assigning that subject to *"organization"* and, of course, less *"organization"* ratings in other subjects (e.g. *"Heather"*) – though still substantial as we saw in our discussion on accuracy. Hence a non-zero excess.

5.4 Overlaps and Redundancy

To explore redundancy, the first exploratory step is to calculate overlaps for all pairs of concepts. The overlap charts of Figure 6 show average overlaps over all relevant items. All overlaps for goal models appear to be bounded by 0.35, while in intention models the overlaps between *"goal"* and *"objective"* (0.58) and *"wants-to"* and *"intends-to"* (0.47) stand out as exactly expected (**O.1**). However, the overlaps between *"goal"* and *"belief"* (0.34) and *"motivates"* with *"is-a-means-to"* (0.3) in goal models, also stand out compared to other pairs. While the latter overlap can be explained by the ease by which the elements belonging to each concept can be confused, the former is less easy to explain.

Calculations of overlaps per participant are also pertinent in this analysis. Of the statistical tests comparing the two benchmark pairs (**O.2**), one rejects the null (relationships - $W = 42, p = 0.0049, r = 0.5\,(large)$) but one fails (entities - $W = 59, p = 0.056$). The latter appears to be due to the high overlap we observed between *goal* and *belief*.

Finally, as per the operationalizations, to specifically study redundancy of a concept we need to collect a low-quantile (over subjects) of the relevant overlaps that the concept exhibits in relation to each other concept. For the analysis, we set a relevance threshold to $\alpha = 0.1$, i.e. an element is relevant for the calculation of an overlap involving two concepts if at least 10% of the total ratings the element receives are for either one of the concepts involved in the pair.

Results for intention models can be found in Table 10. Thus, although the minima are all zero, based on 10th percentiles, for *"goal"* and *"objective"* the corresponding redundancy indices are 0.15 and for the *"wants-to"* and *"intends-to"* 0.03. This means that if we take away 10% of the subjects with the lowest overlap, the next lowest overlap that is observed between the terms of those two pairs is non-zero. As we increase the quantile, other pairs emerge, for which, however, we are less confident that they overlap, in that, in higher quantiles sensitivity increases, i.e., a larger number of non-overlapping subjects needs to be ignored. Thus, the aforementioned strong overlap between *"goal"* and *"belief"* is observed at the 25th percentile in the form of redundancy for both concepts. Notice that, despite this overlap, when confronted with a decision to include in the language either *"objective"* or *"belief"*, designers will not choose the former, as its overlap with *"goal"* is greater – Figure 6. They will likely investigate the specific examples or seek to obtain data from a larger sample.

The above seem to confirm expectations (**O.3**), with the caveat that exploration of various quantiles is needed (Table 10) instead of blindly relying on, e.g. the minimum overlap.

	H_1	Description	Tests	Outcome
Deficit	**D.1.**	Intention models exhibit more deficit than goal models.	$\text{def}_{im} > \text{def}_{gm}$	☆ (with comments)
	D.2.	Deficit-per-participant is greater in intention models.	$\mathbf{def}^{pp}_{V_{im}} >_m \mathbf{def}^{pp}_{V_{gm}}$	☆★
Excess	**E.1.**	Intention models generally exhibit more construct excess than goal models.	$\text{exc}_{im}(v_1) > \text{exc}_{gm}(v_2)$ for corresponding $v_1 \in V_{im}, v_2 \in V_{gm}$ $\mathbf{exc}^{pp}_{im} >_m \mathbf{exc}^{pp}_{gm}$	☆ (with comments) ✪
	E.2.	*"Prevents"* stands-out as excessive.	$\text{exc}_{im}(\text{"prevents"}) > \text{exc}_{[\cdot]}(v)$ for $v \in V_{gm} \cup V_{im}$	☆
	E.3.	*"Prevents"* is more per-participant excessive than *"motivates"*.	$\mathbf{exc}^{pp}_{im}(\text{"prevents"}) >_m \mathbf{exc}^{pp}_{gm}(\text{"motivates"})$	☆★
	E.4.	No strong evidence of excess in all goal model concepts and *"organization"* and *"is-a-means-to"* in intention models.	Low $\mathbf{exc}_{[\cdot]}(v)$ for the mentioned v	☆

Table 9 The experimental hypotheses: deficit and excess
$\mathbf{a} >_m \mathbf{b}$ is interpreted into a one-tail Wilcoxon test for independent samples $\mathbf{a}$ and $\mathbf{b}$
Outcomes: observed qualitatively (☆), and, where statistical test is applicable, passes (★), fails (✪)

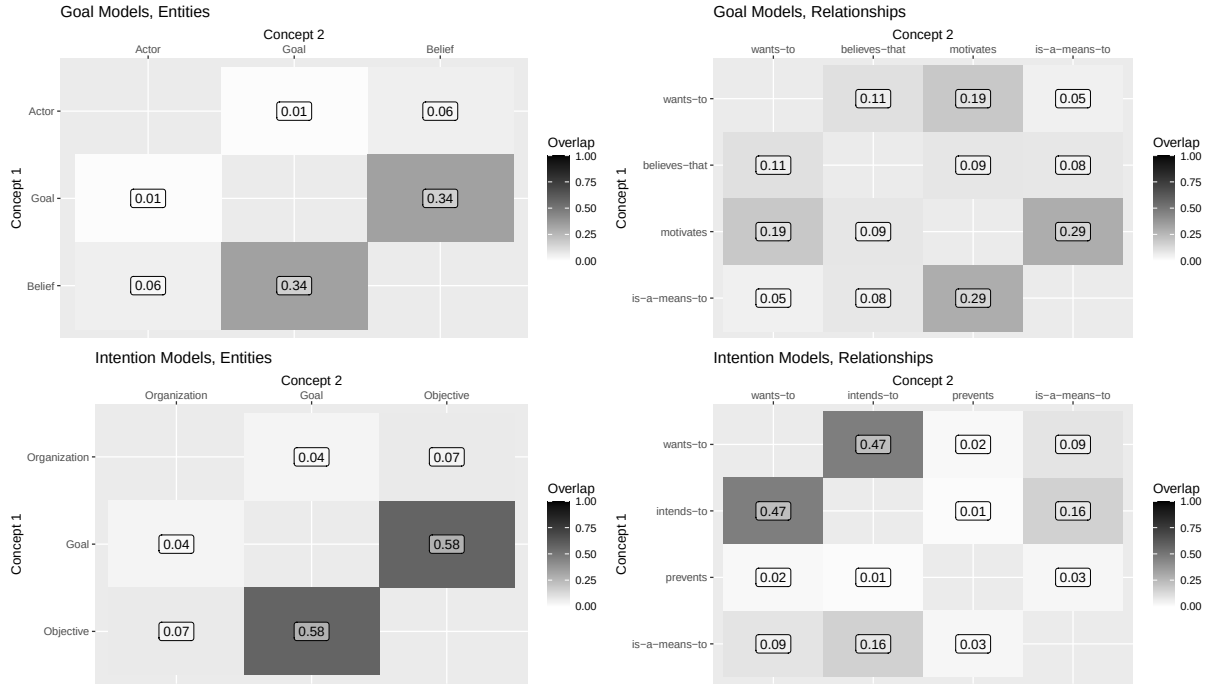

Figure 6 Overlap Chart ($\alpha = 0.1$)

5.5 Self-reported Data

Let us now turn our focus to the self-reported data and their correlation to the observed data, as well as their ability to detect the differences between the two languages.

Deficit. Participants are asked to rate the degree to which they found the vocabulary to be *complete*. The higher the grade the more complete they found the vocabulary to be. Naturally we expect a negative correlation between the observed deficit and the reported completeness (**S.1**). Considering all data, a negative correlation is indeed observed $\tau = -0.25, p = 0.0054$ which however does not pass our discounted alpha threshold in

Language	Concept	min	q10	q25	median
Goal Models	Actor	0.00	0.00	0.00	0.00
	Belief	0.00	0.00	0.21	0.26
	Goal	0.00	0.00	0.21	0.26
	believes-that	0.00	0.00	0.00	0.05
	is-a-means-to	0.00	0.00	0.01	0.29
	motivates	0.00	0.00	0.02	0.29
	wants-to	0.00	0.00	0.02	0.20
Intention Models	Goal	0.00	**0.15**	0.36	0.66
	Objective	0.00	**0.15**	0.36	0.66
	Organization	0.00	0.00	0.00	0.05
	intends-to	0.00	**0.03**	0.07	0.27
	is-a-means-to	0.00	0.00	0.03	0.11
	prevents	0.00	0.00	0.00	0.00
	wants-to	0.00	**0.03**	0.07	0.27

Table 10 Redundancy indexes.

terms of statistical significance. Comparing the self-reported deficit between the two languages, however (**S.4**), turns out to be significant, $W = 539.5, p = 0.0012, r = 0.41$ (*moderate*).

Excess. Participants are asked to rate each term with respect to its relevance. It is expected, therefore, that there would be a negative correlation between the self-reported relevance and the observed excess (**S.2**). This indeed turns out to be the case – considering, again all data, the correlation test is significant, Kendall $\tau = -0.25, p < 0.001$. Comparing the self-reported excess between *"motivates"* and its unfortunate replacement, *"prevents"* (**S.4**), also turns out to be significant: $W = 31.5, p = 0.0019, r = 0.56$ (*large*).

Overlap. Participants are also asked to rate the overlap between pairs of entities and relationships. We again expect these ratings to correlate positively with the measured overlaps. Indeed, the analysis for all data yields a statistically significant positive correlation: $\tau = 0.37, p < 0.001$ (**S.3**). In addition, when restricting self-reported data to the pairs of interest (*"goal"* and *"belief"* vs. *"goal"* and *"objective"*, as well as, *"wants-to"* and *"believes-that"* vs. *"wants-to"* and *"intends-to"*) the overlap ratings of the pairs belonging to goal models are lower than those belonging to intention models (**S.4**) – $W = 74.5, p < 0.001, r = 0.65$ (*large*).

5.6 Test-Retest Reliability

A final concern is that of instrument reliability, which is part of our question whether devising an instrument operationalizing the proposed metrics is at all possible (RQ1). We are specifically testing whether the same participants in different points in time offer responses that are the same or similar. This is known as test-retest reliability [26]. To measure it we invite participants to redo the exact same instrument activities some time after they performed them for the first time. We set that time to be two (2) weeks. From the two sets of data points we calculate Cohen's kappa coefficient [18].

Through test-retest reliability we are interested in measuring the way we elicit responses (e.g. the exact language we use to ask the questions, the intuitiveness and understandability of the on-line instrument, other pragmatic/implementation aspects that may affect reliable measurement) rather than the language itself. Hence, in choosing which participants to invite for a retest, we need to consider that the issues that have been manufactured in intention models are expected to evoke inconsistent responses, confounding our effort to measure reliability of the instrument per se. We, hence, restrict our focus to participants originally assigned to goal models, inviting those who offered qualified responses, and ignore participants assigned to intention models.

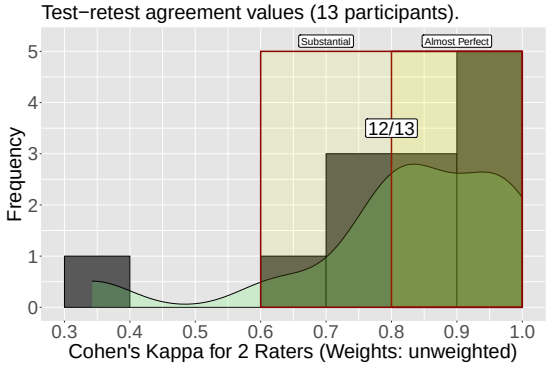

Figure 7 Distribution of values for Cohen's kappa measuring agreement between 2 repeated administrations with the goal model group – two weeks apart.

Of those invited, all thirteen (13) qualify again. The equal number of kappa values that result from the analysis are indeed high: the median value is 0.85 and twelve (12) out of the thirteen (13) values are 0.61 or more, while eight (8) are 0.81 and more. According to Landis and Koch's characterisations [27], 0.61 or higher and 0.81 or higher indicate substantial and almost perfect agreement,

	H_1	Description	Tests	Outcome
Overlap	**O.1.**	Overlaps between *"goal"* and *"objective"* and between *"wants-to"* and *"intends-to"* are substantially higher than all other pairs.	$\text{ov}_{V_{im}}$(*"goal"*, *"objective"*) and $\text{ov}_{V_{im}}$(*"wants-to"*, *"intends-to"*) are higher than all other overlaps	☆ (with comments)
	O.2.	The above pairs overlap (per participant) more in intention models than in goal models after replacing the original concepts.	$\mathbf{ov}^{pp}_{V_{im}}$(*"goal"*, *"objective"*) $>_m$ $\mathbf{ov}^{pp}_{V_{gm}}$(*"goal"*, *"belief"*)	☆✪
			$\mathbf{ov}^{pp}_{V_{im}}$(*"wants-to"*, *"intends-to"*) $>_m$ $\mathbf{ov}^{pp}_{V_{gm}}$(*"wants-to"*, *"believes-that"*)	☆★
	O.3.	Redundancy is higher in the above four concepts compared to all other concepts.	$\mathbf{rdn}_{V_{im}}(v)$ for $r \in$ {*"goal"*, *"objective"*, *"wants-to"*, *"intends-to"*} vs. other v's in V_{gm} and V_{im}	☆
Self-Reported	**S.1.**	There is negative correlation between self-reported completeness and per-participant observed deficit $\mathbf{def}^{pp}_V$ (both languages)	Kendall's τ between the two samples	☆✪
	S.2.	There is a negative correlation between the self-reported relevance and per-participant excess $\mathbf{exc}^{pp}_V(v)$ (all concepts)	Kendall's τ between the two samples	☆★
	S.3.	There is a positive correlation between the self-reported and the per-participant observed overlap $\mathbf{ov}^{pp}_V(v, v')$	Kendall's τ between the two samples	☆★
	S.4.	Self-reported values for deficit (all concepts), excess (targeted data) and overlap (targeted pairs) differ between the two languages.	Deficit (all concepts)	☆★
			Excess (*"motivates"* vs. *"prevents"*)	☆★
			Overlaps (all self-reported data on O.2 pairs)	☆★
Retest	**T.1.**	There is substantial agreement between the two consecutive responses (main test and retest) for most participants.	Observational data (Cohen's kappa)	☆
			Self-reported Data (Krippendorff's alpha)	✗

Table 11 The experimental hypotheses: overlap, redundancy and relationship to self-reported data
a $>_m$ **b** is interpreted into a one-tail Wilcoxon test for independent samples **a** and **b**
Outcomes: observed qualitatively (☆), not observed qualitatively (✗), and, where statistical test is applicable, passes (★), fails (✪)

respectively. We consider these to be strong evidence of test-retest reliability. The distribution of values can be viewed in Figure 7.

A different picture emerges from the parts of the instrument used for eliciting participant opinions on concept overlaps, excess, and vocabulary incompleteness. Given that these are acquired using interval scales, we use Krippendorff's alpha for measuring agreement [17]. The corresponding medians are now 0.64, 0.57, and 0.05 for the questions asking about overlaps for each concept, relevance of each concept (the reverse of excess) and vocabulary incompleteness (measuring deficit), respectively. Accordingly, seven (7), five (5), and one (1) retest participants, respectively, agree with their previous input by an alpha of 0.61 or more. Thus, with the possible exception of self-reported overlaps, the questions eliciting participant opinion cannot be deemed to offer sufficient reliability for direct precise measurement – versus, e.g., a rough preliminary comparison between designs (**S.4**).

6 Study Conclusions and Validity Threats

6.1 Key Findings

Let us now summarize the findings of our experimental study, comment on how they inform our research questions, and discuss the main threats to validity that it faces. The experiment aimed at exploring whether the proposed framework can be applied in practice (**RQ1**), whether it detects

issues that are expected (**RQ2**) without detecting also issues that do not exist (**RQ3**), and how its results compare with ratings generated by the participants themselves (**RQ4**).

The results show that the constructs we devise indeed discover almost all issues that are expected (**RQ2**), either descriptively, i.e., through inspection of the sample data, or statistically/inferentially, i.e., allowing generalization to the entire participant population. For example, the deficit of intention models emerged in the results (hypotheses **D.1.** and **D.2.**), the excessiveness of *"prevents"* was observed (**E.2.** and **E.3.**), and the overlaps introduced in intention models emerged as well (**O.1.** - **O.3.**, though one test fails).

There is further limited evidence of false positives (**RQ3**). The term *"is-a-means-to"* evoked low excess and accurate responses in both languages (e.g., **E.4** and **A.7**) and, with one potential exception (*"goal"* vs. *"belief"*), there are no overlaps other than the ones expected (**O.3**). Equivalence tests, however, would require a much larger sample size. Moreover, more research can be done to investigate whether measures of well-designed concepts and terms are otherwise affected by design flaws in other concepts of the same language. For example, the precision measures of *"is-a-means-to"* are slightly lower in intention models than in goal models – see Table 4. In other words, a healthy term such as *"is-a-means-to"* can be erroneously employed to classify elements for which there is deficit (e.g., *"motivates"*) hurting its own quality measures. More generally, reviewers of the results of an analysis must carefully reason about the indications and their context (descriptions and elements), rather than automatically drawing conclusions, such as assuming that a term such as *"is-a-means-to"* is coarse just because precision may appear to be somewhat lower.

With regards to self-reported data they are found to be consistent with the observational ones (**RQ4**). Firstly, the two sets of data are correlated in the expected way, which constitutes additional validation evidence for the observational measures. At the same time, self-reported data alone can pick up the quality differences between the two languages (**S.4**). It needs to be emphasized, however, that both deficit and excess questions are asked *after* the classification tasks are performed and strongly rely on the participants experience with that task. As such, (a) there is no evidence that they can be used in place of (rather than in addition to) the observational component, in order to make rough evaluations of the language, and (b) like the observational measures, they depend on the choice of descriptions and elements. This is, nevertheless, not the case with self-reported overlaps, which are asked before participants are exposed to any descriptions or elements. Hence, the overlap self-reporting questions can potentially replace the more expensive observational component, though more investigation is needed to confirm this possibility. Finally, while the self-reported data can offer a rough comparison between two languages, the exact numeric value they provide may not be reliable enough to be used as an absolute measure, as the test-retest analysis reveals. This does not seem to be the case with the observational component, which exhibited good test-retest quality (**T.1**).

Finally, we can conclude that generating meaningful operationalizations of the metrics is possible (**RQ1**) in that it allows the development of instruments that both yield meaningful data, as observed above, and show high-levels of reliability (**T.1**).

6.2 Validity Threats and Limitations

Like any empirical study, ours is also exposed to validity threats. These validity threats concern not only the present study but any practical application of the proposed framework. Hence, we discuss the most important ones – external, internal and construct validity – aiming at commenting both on our study and on possible limitations and pitfalls of the proposed framework.

In terms of *external validity*, our conclusions are influenced by the choice of three kinds of samples: the participants, the languages, and the descriptions and domain elements. With regards to participants, the framework suggests samples from the population of people who will produce or consume models using the language in question. For the specific experiment, the languages in question are primarily meant to be used for requirements analysis. We, hence, assume that the intended population is business and requirements analysts. We, further, make the assumption that these roles require an undergraduate degree

in information technology (IT). In opposition to this decision, it can be argued that business analysts or other classes of stakeholders who would use languages with concepts relating to the ones we investigated may not need to have a technical background. Nevertheless, we could not find any empirical evidence confirming or rejecting this assertion, despite its intuitiveness. Regardless, restricting the sampled population, at worst restricts the generalization class to people with IT background. This does not interfere with our basic study objectives which are to offer a demonstration of the validity of the proposed framework, assuming a language and some audience for it. Future work could investigate the role of individual differences in, e.g., the sensitivity or reliability of instruments derived from the framework for different vocabularies.

Furthermore, the results are affected by the choice of the languages under comparison. Firstly, we clarify that, by design of the framework, findings of a study using one language are specific to that language and not amenable to systematic generalization to other languages. Considering the languages used in the experiment, intention models have issues that are unusual in their obviousness, leaving one wondering if a more realistic comparison, between, e.g., two existing languages, would offer results of similar clarity. However, the goal of this study is to establish the possibility of detection of issues via comparison with an external criterion, which, in our case, is the assumed obviousness of the issues with one of the two languages. This criterion would not have been reliable if the quality differences were not intuitively clear, but, rather, a subject of debate or relative viewpoint. For example, if we compared two existing and actively supported languages, for which, hence, there is no consensus as to which one is better designed, there would be no commonly accepted criterion to compare our results with. Nevertheless, while this design choice serves the purpose of this introductory study, it makes our results difficult to generalize to cases in which the differences between the languages under comparison are more nuanced. It is possible, for example, that the metrics become too sensitive to extraneous variables (e.g., expression and rater sampling) when the language issues are more subtle, casting the metrics "noisy" and unreliable for comparisons. Hence, the question of accuracy and precision of the proposed metrics when quality differences are less conspicuous seems to be a matter to be explored through follow-up applications and experiments.

Moreover, the choice of elements to be classified and the descriptions from which they are taken, may affect generalizability. A different set of elements or domains could arguably offer us a different view of how goal models and intention models differ. For instance, simply using, e.g., the banking or aviation domains instead of retail and travel that we used, could have an effect on the result. Characteristics of the description texts including, e.g., their format and structure, their formality and specialization level, or their length may affect the rating process. As we mentioned earlier, application of the framework relies on the evaluators properly identifying descriptions and elements that are representative of the domains of interest and also complete with regards to the concerns they wish to model. The exact methodology for identifying, developing, and evaluating such material, as well as the ways by which bias can be inserted in it requires further investigation, likely focused on methodological aspects. One approach, for example, is to complement the formative evaluations performed by the designers with summative ones performed by independent evaluators, each group creating their own evaluation set (descriptions and elements). If the suspicion of biased sampling is stronger, such evaluation could even be double-blind: as designers are not aware of the evaluation set when they design their language, independent evaluators are not aware of the language concepts as they sample the test set, either.

An additional concern is that of scalability of the metrics to languages with more concepts, from both an external validity and a practicality viewpoint. Factors that are affected by larger languages include substantially heavier training requirements, longer and more cumbersome instruments (e.g., inventory questionnaires with long lists of choices), and larger samples of required expressions and elements to offer adequate coverage. Strategies for addressing these challenges seem to depend on the application at hand. A possible approach is to perform targeted investigations to subsets of the language. For example, a language like Archimate [10] includes a total of several tens of elements and relationships

for modeling enterprise architecture (EA). Rather than evaluating them all at once, one can observe that they are thematically organized with respect to the exact EA layer they are supposed to model (strategy, business, technology, application, motivation, etc.). Experimenters may start by tackling each layer separately, by, e.g., collecting expressions and elements relating to a layer (e.g., strategy) and restricting data collection instruments to concepts included in that layer (the strategy concepts). In other cases, evaluators may want to study an arbitrary subset of concepts against a suspected indication, such as measuring the construct excess or precision of a specific term. Such a practice requires careful explication of assumptions and validity threats. For example, when arbitrarily focusing on a subset of terms, construct deficit indications can be due to exclusion of relevant terms.

As a last comment on external validity, it is worth noting that Peira can be seen as a platform suitable for systematically performing replication studies. For example, the experiment we conducted in this paper is easily replicable [24]. Replication researchers may perform the exact same procedures with a different participant sample, or, for a more indirect replication, may choose to e.g. use different descriptions and domain elements. Peira facilitates replication through the explication of recommended rating procedures and metrics, which makes them unambiguously reproducible.

Turning, further, to *internal validity*, a possible concern is the grammatical properties of the elements vis-à-vis their classification. Consider the elements *"Heather"*, *"Introduce more products"*, and *"Inflation is rising"* under the goal modeling language we discussed. Participants may classify these elements under *actor*, *goal*, and *belief*, respectively, to the satisfaction of the designers. However, it can be the case that participants, informed by any amount of training, respond to the grammatical format of the element rather than its meaning: if it is a noun then it is an actor, if it is in imperative mood it is a goal, and if it is in indicative mood it is a belief. This way, a conceptualization that performs well in our measurements, may fail in the real world where concept instances may have not been extracted and articulated in the way the participants were trained to. Again, here, the onus is on the evaluators to foresee and rule out such effects. This may come in a form of test elements and alternate grammatical formats. For instance, elements *"Storefront"* and *"The desired state of having introduced more products"* could be introduced to test if participants assign under *"actor"* just any noun or are confused by presentations of *"goal"* that are more unusual in daily discourse.

Moreover, training quality and duration is expected to affect the results. A perfectly designed language may score low because its guiding and training material is poor, or because it simply takes a lot of exposure and practice to learn. We have found training to be a recurring internal validity issue in the empirical evaluation and comparison of languages and models [28–30] and obvious solutions may be expensive and complicated. For example, an approach to address biases emerging in cases in which evaluators are proponents of one of the languages under comparison, is to allow for a third disinterested party to develop the training material and/or conduct the training. The effect of training can otherwise be evaluated by holding all other factors constant and varying training variables, either as separate experiments or in the context of complex factorial designs, where training is treated as a covariate.

Further, the choice of domain descriptions and domain elements is a likely source of internal validity concerns. For example, construct excess may be the result of omission of key domain elements rather than a problem with the language, while incomprehensible or ambiguous descriptions may be the source of misclassifications not warranted by the language design per se. As discussed, different strategies for systematically developing descriptions need to be explored in the future. An additional question is whether the descriptions should be manufactured by the evaluators or sampled from the domain and presented verbatim to participants. The latter practice appears to lift the risk of insertion of bias by the evaluators. However valid deficit measures rely on the presence of descriptions that cover all relevant aspects the language is meant to model. Thus, unless the study excludes construct deficit investigation, evaluators need to consciously sample descriptions to ensure such completeness.

An additional comment can be made with regards to *construct validity*, that is, what is really

measured in our experiment and the framework in general. In practice, the framework evaluates a package that consists of (a) a choice of concepts, (b) a set of signifiers for representing those concepts and, importantly, (c) a set of materials and/or training activities for learning the language and its use. Low quality measures can be, hence, attributed to issues in any of the three components. For example, lack of accuracy for a concept may be because of the wrong choice of term to represent it, because the concept itself is too vague or foreign to the daily experience of domain actors, or because the training guides and lessons did not explain it properly or used bad examples. As above, replications may allow discerning the exact issue. For example, if the accuracy issue remains despite attempting different terms or definitions and training examples, we may tend to believe that the choice of concept is sub-optimal.

The key conclusion relating to the majority of the above concerns seems to be that while a single experiment may offer us a preliminary idea of the quality of a language, thorough evaluation may require a family of such experiments.

As a final comment, it is worth discussing how *construct overload*, one of the four quality issues of Wand and Weber's framework [19], can be empirically measured. Recall that construct overload refers to cases in which one term is used to represent more than one concept. The proposed framework does not introduce a concrete metric for it, in that the construct is not available for direct observation through a single round of user classifications. Rather, we propose that overload is assessed through repeated studies in which the language is altered and re-evaluated.

Consider the extreme example of a language with one and only concept, termed *"concept"*. The language is bound to produce very good qualities in terms of accuracy, deficit, excess, or redundancy, as most raters will agree that most subjects can be trivially classified under *"concept"*. We may however suspect that it has high degrees of overload: the term *"concept"* is probably used to refer to a variety of more specialized concepts that we would be interested in referring to with different terms, increasing the expressiveness of the language. We subsequently increase the granularity [31] of the language via replacing this high-level term with more specialized ones. If we subsequently observe that none of the metrics of interest are substantially worsening, we implicitly show that the original language suffered from construct overload. It follows that overload can then be detected methodologically: if we refine a language into one that performs well in all other quality aspects, we implicitly indicate the presence of remediable construct overload in the original language.

7 Related Work

Evaluation of modeling language quality has been an active area of research within the field of conceptual modeling. Several efforts for organizing the dimensions along which such quality can be characterized have been introduced, including SEQUAL [16], as well as the Conceptual Modeling Quality Framework (CMQF) [32] which aims at combining ideas from two earlier such frameworks by Wand and Weber [33] and by Lindland et al. [34]. The central empirical constructs we introduced here reflect SEQUAL's quality dimensions *domain appropriateness* and *comprehensibility appropriateness*. In his own framework concerned primarily with visual representation, Moody [35] uses the notions of *semantic transparency* and *semiotic clarity* to describe whether visual signifiers evoke the intended concepts. These frameworks offer a comprehensive view of aspects of language quality that are important, and are often used for analytical evaluation of modeling notations (e.g., [36]). However, development of empirical (e.g., experimental) constructs and procedures, such as the ones we attempt in this paper, is important both for systematically and reliably measuring such quality aspects and for attaining a shared understanding of what such quality dimensions really mean. Work with a similar emphasis on the measurement aspect, but with a stronger focus on the organization of the overall empirical procedure has been reported by Bork et al. [37, 38]

In the area of empirical conceptual modeling, substantial effort has been dedicated towards assessing model understandability of various notations. An extensive survey of relevant studies as a means to an introduction to the problematic of measuring understandability is offered by Houy et al. [39]. Many such comprehensibility or other quality assessment studies have been conducted

in the area of goal modeling from where our example languages have been adopted [40–44]. The notion of intuitive comprehensibility, for example, has been intensively applied by Liaskos et al. in various studies [28–30, 45], aimed at measuring the level to which visual signifiers naturally (e.g., without specific training) evoke the meaning of the signified concept, through observing inferences participants perform with the models. Compared to these works, in the work we presented, rather than merely detecting such misalignments, we attempt a more refined characterization of the nature and quality of misalignment between concept and its visual/verbal signifier, and offer a toolset for systematically performing such analyses.

Procedures similar to qualitative coding, such as annotation of text, have been introduced in the area of ontology engineering as well [46, 47], where application of notions of inter-rater agreement [17, 18] have also been proposed [48]. Our work is also relevant to ontology learning techniques whereby processes such as term extraction are utilized [49] for identifying concept-describing terms from domain text. Our approach shares with these efforts the principle of term choices being grounded on domain information, albeit it presumes a design stage in which a set of candidate terms has already been defined.

The use of formal ontology specifications has been understood to be an additional way by which the sharedness of ontological commitments is promoted [7]. This is done through, e.g, formulation of properties of language terms that follow from their intended meaning, i.e., are consequences of the commitment. In addition, upper-level ontologies have been proposed as one of the ways to analytically identify issues with a language metamodel [50]. Empirical analysis, such as the one we promote with the proposed framework, is not meant to substitute but to complement other tools that language evaluators have at their disposal for ensuring optimal language design.

Finally, it is worth looking at the relationship of our framework with representational measurement theory [51–53]. The main concern in that context is the interpretation of empirically observed qualitative relations into numerical scales. Thus, construction of a scale requires the development of a qualitative relational structure and its axioms (e.g., is it transitive? is it reflexive? etc.), a mapping of that structure to an appropriate axiom-satisfying numerical relational structure, and the identification of allowable transformations of the latter. On one hand, our core framework (the set of abstract metrics) deliberately avoids proposing concrete scales. It rather describes the principles under which such scales can be constructed in terms of counting participant rating events. Thus, it is up to instantiations of the framework to attempt such analyses when warranted and meaningful. On the other hand, however, proposing general axiomatizations to be part or Peira's key metrics can prove to be a useful future addition to the framework, in that it offers a foundation for systematically developing concrete metrics.

8 Summary and Future Work

We presented Peira, a framework for empirically evaluating modeling language ontology and vocabulary qualities. Peira achieves this through first observing how samples of raters from the intended language user population use the concept representations to model relevant domain phenomena and then using such data for calculating statistical measures indicating a variety of types of issues. The framework defines such types of issues in an abstract sense allowing adopters to define concrete metrics that fit their specific interests, analysis plans, and data collection instruments. An experimental study comparing two languages, one that is established and widely studied against another where specific issues have been introduced, reveals that the development of appropriate instruments and operationalizations is possible and that it allows for detecting the most important issues even with few and conveniently sampled participants.

There are different directions towards which this work can be extended. One is the conduct of more studies in order to better understand the measurement accuracy and the distribution of metric values when evaluating different languages, the role of participant and description/element sampling, or the role of training methods, modes and durations. Such studies can involve both artificial languages, so that a reliable external criterion can be used for validating the metric results, and real languages, so that the behavior of the metrics in the presence of subtle issues is

explored. Moreover, the more studies of the latter kind are conducted and published using the same constructs, the more it will be possible to establish publicly available populations of measures (test norms) such that individual values can be characterized with respect to their ranking in such populations. Development of such norms are beneficial in the long term, as they will allow evaluations of designs without the need for comparative analyses.

Alternative data collection instruments can also be attempted, such as constructive term-centered ones in which participants build extensions for each term in the language – versus the subject-centered approach we applied in our study in which each subject is classified to a term – or ones that allow classification into visual signifiers in addition to just terms. In parallel, work can be dedicated towards the development of more reliable self-reporting instruments that can be used as a companion or a cost-effective alternative to the observational metrics that we presented. Establishing reliability is the first step towards standardization of such instruments, which, in turn, brings the benefit of systematic data acquisition, as well as comparability with similar studies and, again, test norms. Finally, a measurement-theoretic analysis of the proposed metrics – both observational and self-reporting – appears to be an important step to both paving the way for systematic definition of operationalizations and, through developing appropriate axiomatizations within such work, better defining the corresponding quality notions (overload, semantic overlap, etc.).

We believe that progress in these fronts would greatly assist the design of intuitive and comprehensibility and domain appropriate conceptual modeling languages, which are, in turn, crucial to the analysis, design, and development of software-intensive systems.

References

[1] Fettke, P.: How Conceptual Modeling Is Used. In: Communications of the Association for Information Systems, vol. 25 (2009). https://doi.org/10.17705/1CAIS.02543 . https://doi.org/10.17705/1CAIS.02543

[2] 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) https://doi.org/10.1016/j.datak.2005.07.007

[3] Olivé, A.: Conceptual Modeling of Information Systems. Springer, Berlin; Heidelberg (2007)

[4] Alexander T. Borgida Vinay K. Chaudhri, P.G., Yu, E.S. (eds.): Conceptual Modeling: Foundations and Applications. Springer, Berlin; Heidelberg (2009)

[5] Mylopoulos, J., Borgida, A., Jarke, M., Koubarakis, M.: Telos: Representing Knowledge about Information Systems. ACM Transactions on Information Systems **8**(4), 325–362 (1990) https://doi.org/10.1145/102675.102676

[6] Karagiannis, D., Khun, H.: Metamodelling Platforms. In: Proceedings of the Third International Conference on E-commerce and Web Technology (EC-Web 2002), pp. 182–197 (2002)

[7] Guarino, N., Oberle, D., Staab, S.: What Is an Ontology? In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 1–17 (2009). https://doi.org/10.1007/978-3-540-92673-3_0

[8] Object Management Group: OMG Unified Modeling Language (OMG UML) – Version 2.5.1. (2017). Object Management Group. https://www.omg.org/spec/UML/2.5.1/PDF

[9] Dalpiaz, F., Franch, X., Horkoff, J.: iStar 2.0 Language Guide. The Computing Research Repository (CoRR) **abs/1605.0** (2016) arXiv:1605.07767

[10] The Open Group: ArchiMate® 3.1 Specification. Technical report (2019)

[11] What's New in ArchiMate 2.0? https://blog.opengroup.org/2012/01/31/whats-new-in-archimate-2-0/. [Online;

accessed 25-November-2023] (2012)

[12] What's new in the ArchiMate 3.0 modeling language? https://blog.opengroup.org/2016/06/14/whats-new-in-archimate-3-0/. [Online; accessed 25-November-2023] (2016)

[13] ArchiMate® 3.1 Specification: The New Version of the Standard. https://blog.opengroup.org/2012/01/31/whats-new-in-archimate-2-0/. [Online; accessed 25-November-2023] (2019)

[14] Liaskos, S., Mylopoulos, J., Khan, S.M.: Empirically Evaluating the Semantic Qualities of Language Vocabularies. In: Ghose, A.K., Horkoff, J., Souza, V.E.S., Parsons, J., Evermann, J. (eds.) Proceedings of the 40th International Conference on Conceptual Modeling (ER 2021). Lecture Notes in Computer Science, vol. 13011, pp. 330–344. Springer, Berlin, Heidelberg (2021). https://doi.org/10.1007/978-3-030-89022-3_26 . https://doi.org/10.1007/978-3-030-89022-3%5C_26

[15] Dickover, M.E., McGowan, C.L., Ross, D.T.: Software Design Using: SADT. In: Proceedings of the 1977 Annual Conference of the ACM. ACM '77, pp. 125–133. Association for Computing Machinery, New York, NY, USA (1977). https://doi.org/10.1145/800179.810192 . https://doi.org/10.1145/800179.810192

[16] Krogstie, J.: Model-Based Development and Evolution of Information Systems. Springer, Berlin; Heidelberg (2012)

[17] Krippendorff, K.: Content Analysis: An Introduction to It Methodology. SAGE, Thousand Oaks; London; New Delhi (2004)

[18] Kilem L Gwet: Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC, Gaithersburg (2014)

[19] Wand, Y., Weber, R.: On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal **3**(4), 217–237 (1993)

[20] Stoet, G.: PsyToolkit: A software package for programming psychological experiments using Linux. Behavior Research Methods **42**(4), 1096–1104 (2010) https://doi.org/10.3758/BRM.42.4.1096

[21] Stoet, G.: PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. Teaching of Psychology **44**(1), 24–31 (2017) https://doi.org/10.1177/0098628316677643

[22] Prolific. https://www.prolific.co/ (2022)

[23] Peer, E., Rothschild, D., Gordon, A., Evernden, Z., Damer, E.: Data quality of platforms and panels for online behavioral research. Behavior Research Methods **54**(4), 1643–1662 (2022) https://doi.org/10.3758/s13428-021-01694-3

[24] Liaskos, S., Zarbaf, S.: Replication Data For: Empirically Evaluating Modeling Language Ontologies: the Peira Framework. https://doi.org/10.5683/SP3/O1E4PL.

[25] Dikta, G., Scheer, M.: Bootstrap Methods With Applications in R, 1st edn. Springer, Berlin; Heidelberg (2021). https://doi.org/10.1007/978-3-030-73480-0

[26] Rosnow, R.L., Rosenthal, R.: Beginning Behavioral Research: A Conceptual Primer, 6th edn. Pearson Prentice Hall, NJ, USA (2008)

[27] Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics **33**(1), 159–174 (1977) https://doi.org/10.2307/2529310

[28] Alothman, N., Zhian, M., Liaskos, S.: User Perception of Numeric Contribution Semantics for Goal Models: an Exploratory Experiment. In: Proceedings of the 36th International Conference on Conceptual Modeling (ER 2017), Xi'an, China, pp. 451–465 (2017). http://www.yorku.ca/liaskos/Docs/ER17.pdf

[29] Liaskos, S., Ronse, A., Zhian, M.: Assessing the Intuitiveness of Qualitative Contribution Relationships in Goal Models: an

Exploratory Experiment. In: Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'17), Toronto, Canada, pp. 466–471 (2017). http://www.yorku.ca/liaskos/Docs/ESEM17.pdf

[30] Liaskos, S., Dundjerovic, T., Gabriel, G.: Comparing Alternative Goal Model Visualizations for Decision Making: an Exploratory Experiment. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC'18), Pau, France, pp. 1272–1281 (2018). http://www.yorku.ca/liaskos/Papers/SAC2018/Visualizations/SAC2018.pdf

[31] Henderson-Sellers, B., Gonzalez-Perez, C.: Granularity in Conceptual Modelling: Application to Metamodels. In: Proceedings of the 29th International Conference on Conceptual Modeling (ER 2010), Vancouver, BC, Canada, pp. 219–232 (2010)

[32] Nelson, H.J., Poels, G., Genero, M., Piattini, M.: A conceptual modeling quality framework. Software Quality Journal (20), 201–228 (2012)

[33] Wand, Y., Weber, R.: Toward a Theory of the Deep Structure of Information Systems. In: Proceedings of the Conference on Information Systems (ICIS 1990), pp. 61–71 (1990)

[34] Lindland, O.I., Sindre, G., Solvberg, A.: Understanding quality in conceptual modeling. IEEE Software **11**(2), 42–49 (1994)

[35] Moody, D.L.: The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. IEEE Transactions on Software Engineering **35**(6), 756–779 (2009) https://doi.org/10.1109/TSE.2009.67

[36] Moody, D.L., Heymans, P., Matulevičius, R.: Visual syntax does matter: improving the cognitive effectiveness of the i* visual notation. Requirements Engineering **15**(2), 141–175 (2010)

[37] Bork, D., Roelens, B.: A technique for evaluating and improving the semantic transparency of modeling language notations. Software and Systems Modeling **20**(4), 939–963 (2021) https://doi.org/10.1007/s10270-021-00895-w

[38] Bork, D., Karagiannis, D., Pittl, B.: How are Metamodels Specified in Practice? Empirical Insights and Recommendations. In: Proceedigns of the 24th Americas Conference on Information Systems (AMCIS'18) (2018)

[39] Houy, C., Fettke, P., Loos, P.: Understanding understandability of conceptual models - What are we actually talking about? In: Proceedings of the 31st International Conference on Conceptual Modeling (ER 2012), vol. LNCS 7532, pp. 64–77 (2012)

[40] Caire, P., Genon, N., Heymans, P., Moody, D.L.: Visual notation design 2.0: Towards user comprehensible requirements engineering notations. In: Proceedings of the 21st IEEE International Requirements Engineering Conference (RE'13), Rio de Janeiro, Brasil, pp. 115–124 (2013)

[41] Estrada, H., Rebollar, A.M., Pastor, O., Mylopoulos, J.: An Empirical Evaluation of the i* Framework in a Model-Based Software Generation Environment. In: Proceedings of the 18th International Conference on Advanced Information Systems Engineering (CAiSE'06), pp. 513–527. Springer, Luxembourg, Luxembourg (2006)

[42] Hadar, I., Reinhartz-Berger, I., Kuflik, T., Perini, A., Ricca, F., Susi, A.: Comparing the comprehensibility of requirements models expressed in Use Case and Tropos: Results from a family of experiments. Information and Software Technology **55**(10), 1823–1843 (2013)

[43] Horkoff, J., Yu, E.: Finding solutions in goal models: an interactive backward reasoning approach. In: Proceedings of the 29th International Conference on Conceptual Modeling (ER'10). ER'10, pp. 59–75, Vancouver, Canada (2010)

[44] Santos, M., Gralha, C., Goulão, M., Araújo,

J.: Increasing the Semantic Transparency of the KAOS Goal Model Concrete Syntax. In: Proceedings of the 37th International Conference on Conceptual Modeling (ER'18), Xi'an, China, pp. 424–439 (2018)

[45] Liaskos, S., Tambosi, W.: Factors Affecting Comprehension of Contribution Links in Goal Models: An Experiment. In: Proceedings of the 38th International Conference on Conceptual Modeling (ER'19), Salvador, Brazil, pp. 525–539 (2019)

[46] Cimiano, P., Mädche, A., Staab, S., Völker, J.: Ontology Learning. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 245–267. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_11 . https://doi.org/10.1007/978-3-540-92673-3_11

[47] Wong, W., Liu, W., Bennamoun, M.: Ontology Learning from Text: A Look Back and into the Future. ACM Computing Surveys **44**(4) (2012) https://doi.org/10.1145/2333112.2333115

[48] Obrst, L., Ceusters, W., Mani, I., Ray, S., Smith, B.: The Evaluation of Ontologies. In: Baker, C.J.O., Cheung, K.-H. (eds.) Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, pp. 139–158. Springer, Boston, MA (2007). https://doi.org/10.1007/978-0-387-48438-9_8 . https://doi.org/10.1007/978-0-387-48438-9_8

[49] Medelyan, O., Witten, I.H.: Thesaurus-based index term extraction for agricultural documents. In: Proceedings of 2005 EFITA/WCCA Joint Congress on IT in Agriculture, pp. 1122–1129. EFITA/WICCA, Conference held at Vila Real, Portugal (2005). https://hdl.handle.net/10289/8101

[50] Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. PhD thesis, University of Twente (2005)

[51] Krantz, D.H., Luce, D.R., Suppes, P., Tversky, A.: Foundations of Measurement Volume I: Additive and Polynomial Representations. Academic Press, ??? (1971)

[52] Narens, L.: Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science. Lawrence Erlbaum Associates, Inc., ??? (2007)

[53] Hand, D.J.: Statistics and the Theory of Measurement. Journal of the Royal Statistical Society. Series A (Statistics in Society) **159**(3), 445 (1996) https://doi.org/10.2307/2983326