The EmVoc framework for empirically evaluating modeling language vocabulary qualities: a pilot evaluation study

Sotirios Liaskos, Saba Zarbaf

Abstract—Conceptual modeling is central for a variety of activities surrounding the planning, design, development and maintenance of software-intensive systems. To allow development of conceptual models that are understood in the same way by different people, conceptual modeling languages are developed which contain concepts and rules that dictate how correct models can be built according to the language. A key component of a modeling language is a set of concepts that modelers must use to describe world phenomena. Once the concepts are chosen, a visual and/or linguistic vocabulary is adopted for representing the concepts. Both the choices of concepts and the vocabulary used to represent them, however, may affect the quality of the language under consideration. Some choices may match the intentions of the language or may allow for a shared understanding of the concepts better than other choices. We present EmVoc, a framework for empirically measuring the appropriateness of conceptual modeling language vocabularies based on observing how language user sample classify domain content under concepts of the language. The framework is based on a set of abstract empirical constructs that show how such empirical observations can be analyzed in order to detect vocabulary issues of various kinds. The constructs can be operationalized into concrete measures based on the specific data collection instrument or interests of the study. As a first evaluation of our framework we apply it to compare an existing language with an artificial one that is manufactured to exhibit specific issues. We then test if the metrics indeed detect these issues and only. In this paper, we present the complete experimental design and report on the results of a pilot administration. The complete study will be reported in a future version of this paper.

Index Terms—Conceptual Modeling, Empirical Methods

1 Introduction

Developing conceptual models is known to be vital for the effective planning, development and maintenance of software-intensive systems [1], [2], [3], [4], [5]. They facilitate communication of a variety of aspects of such systems and their development process ranging from the system requirements, architecture and design to the business process and enterprize structures in which such systems are embedded.

A fundamental requirement for a successful conceptual model is that it can be understood by different people in the same way. To achieve that, *conceptual modeling languages* are developed and used. Such languages include a number of components and rules that dictate how modelers are supposed to develop correct models that abide by the language and how readers of the models are supposed to understand them so that communication takes place accurately. Such components and rules include a syntax, a visual notation, semantics, as well as accompanying model, development analysis methods and tools and are usually described in training and guiding documentation and material.

At the core of every conceptual modeling language lies a *conceptualization*, that is, a set of concepts that users of the language (modelers and readers) use to name phenomena in the world that are interesting to communicate. Thus, languages for modeling world phenomena as state transitions, define conceptualizations containing concepts such as *state*, *transition*, and *guard condition* (e.g., [6]) while languages for representing stakeholder intentions for requirements engineering use concepts such as *goal*, *task* and *actor* (e.g., [7]). Thus, when designing a new language or revising an existing one, language designers must decide what concepts the language should include and with what natural language terms (e.g., "state", "goal", etc.) these concepts should be referred to.

Clearly, however, given a specific kind of world phenomena that language designers want their language to used for, one set of concepts may be a better choice than another set of concepts. Likewise, for a chosen concept one choice of term may describe it better than another. Traditionally, such choices seem to have largely been a matter of authoritative choice by the language designers based on their experience in the domain and in language design. However it can also be claimed that evaluation of candidate conceptualizations and terms to describe them can be assisted by empirical measurement of whether prospective users of the language correctly associate language concepts with expressions of states of the world that such concepts are meant to represent. If such empirical measurement were possible, simple empirical tests could be conducted to evaluate whether the designers and the intended users of the language are in

- *Sotirios Liaskos is with the School of Information Technology, York University, Toronto, Canada.*
  E-mail: liaskos@yorku.ca
  Web: http://www.yorku.ca/liaskos
- *Saba Zarbaf is with the Department of Electrical Engineering and Computer Science, York University, Toronto, Canada.*
  E-mail: sabazb@cse.yorku.ca

First posted June 10th, 2022; revised XXX,XXX

agreement on what the concepts of a proposed conceptualizations mean, and whether these concepts can completely and concisely represent the phenomena of interest.

In this paper, we present and evaluate a framework for empirical measurement of conceptualization qualities of conceptual modeling languages. The framework is based on the adoption and extension of a system for characterizing language issues and the introduction of a set of abstract empirical *measurement constructs* [8] that show how such issues can be measured from empirical data. The constructs are based on having samples of intended language users classify world-describing expressions under the concepts of the proposed conceptualization, and measuring their in-between agreement, as well as their agreement with the designers' authoritative classifications. Depending on the specific evaluation needs, the constructs can then can be operationalized in ways that fit the specific data collection method.

For evaluating the proposal we conduct an experiment in which we consider two languages: an established conceptual modeling language assumed to have limited or known issues and a artificial language developed for the purpose of the experiment, in which specific alignment issues are deliberately introduced. The main goal of the experiment is to observe if the issues that are manufactured in the second language turn out to be detected by the proposed metrics and operationalizations thereof. In addition, we aim at exploring if attitudinal data collected from the participants, in which they describe their opinion about the appropriateness of the conceptualization correlate with the observations on how the participants actually decided to associate concepts with given world descriptions. The paper extends our earlier paper on the matter [9] with more technical details on the operationalization aspect and the above experimental evaluation that is based on real rather than simulated data.

The paper is organized as follows. In Section 2 we present the key ideas and concepts pertaining to empirical evaluation of conceptualizations using measures of alignment between concepts and world descriptions. In Section 3 we present our constructs and method. Then in Section 4 we present our experimental design, including the research questions, operationalizations and hypotheses we investigate. In 5 we present the results of the pilot study and we conclude with next steps in Section 6.

2 Background

2.1 Languages, Conceptualizations and Vocabularies

One of the primary property of an effective conceptual model is that everyone understands it in the same way. To allow the development of models that evoke a common understanding among those that develop and read them, *conceptual modeling languages* are used. Such languages contain a set of rules and components that must be utilized for the development of models, if the latter are to be presented as valid models of the language, inviting readers to understand them on the basis of that language.

At the core of a conceptual modeling language lies a *conceptualization* i.e. a set of concepts to be used for characterizing phenomena in the domain of interest. To develop the notion of 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 [10].

Firstly, an *extensional relation* is simply a set of ordered n-tuples constructed with elements from D. 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. Hence, D = {(isOn), (isOff), (powerButtonPressed), ...}. An extensional n-ary relation may be a set of pairs or n-tuples of elements from D that we are interested in representing. For example the set of triples {(isOn, powerButtonPressed, isOff), (isOff, powerButtonPressed, isOn), ...} may be a relation that signifies valid transitions to the state specified in the third place of the triple, when the event signified in the second place of the triple takes place when the system is in state signified by the first place of the triple.

Following always the presentation by Guarino et al., concepts represent our ability to identify relations over D under different states of the system S, i.e., under different worlds in W. Specifically a **concept** is a total function $\rho^n : W \mapsto 2^{D^n}$ from worlds to all possible extensional n-ary relations on D. Further, a **conceptualization** is a set of concepts $\mathcal{R}$ on the domain space $< D, W >$.

Thus, in the aforementioned printer example, the concept *event* maps possible, e.g., versions of the system to possible extensional unary relations (sets of elements) from D. Hence in a word w_1, $\rho(event)$ = {(powerButtonPressed), (cancelButtonPressed), ... } and in a world w_2, say, after the firmware is updated, a different extension is mapped to the concept *event* that, e.g., includes (jamDetected). A conceptualization is a set of such concepts comprising a language, e.g. in a language for state transitions it includes {*event, state , state transition, guard condition, ...* }

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 [6]. 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 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.

2.2 Ontological Commitment and Conceptualization-Vocabulary Alignment

Once the vocabulary is defined, it is important to ensure that the 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 with elements or n-tuples from D. Obviously, we want the language to allow for each term $v \in V$ only models 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, again, the part of UML utilized for the production of state diagrams. As we saw, 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. Thus, for a domain $D = \{$(isOn), (isOff), (powerButtonPressed)$\}$, only the last element (powerButtonPressed) is allowed to be in the extension of "trigger" if we are to abide by the obvious ontological commitment of the term to the concept *trigger*. Likewise for the same domain, only the first two elements can be in the extension of term "state" if we are, again, to be faithful to the ontological commitment of said term to the concept *state*.

An ontological commitment can be enforced through introducing axioms in the language, such as meaning postulates [10], 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, as demonstrated above, language users often need to rely on the meaning the vocabulary term in the natural language from where it is borrowed (e.g., English) plus informal definitions and examples in language specifications. guides and tutorials.

Thus, ensuring that the ontological commitment of a language is shared – meaning that users of the language associate signifiers to concepts in the exact way that the language designers intended them to – relies on a choice of signifiers such that the latter evoke the right concept by users based on the user's knowledge of the term in the natural language from where it is borrowed 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', all else being equal.

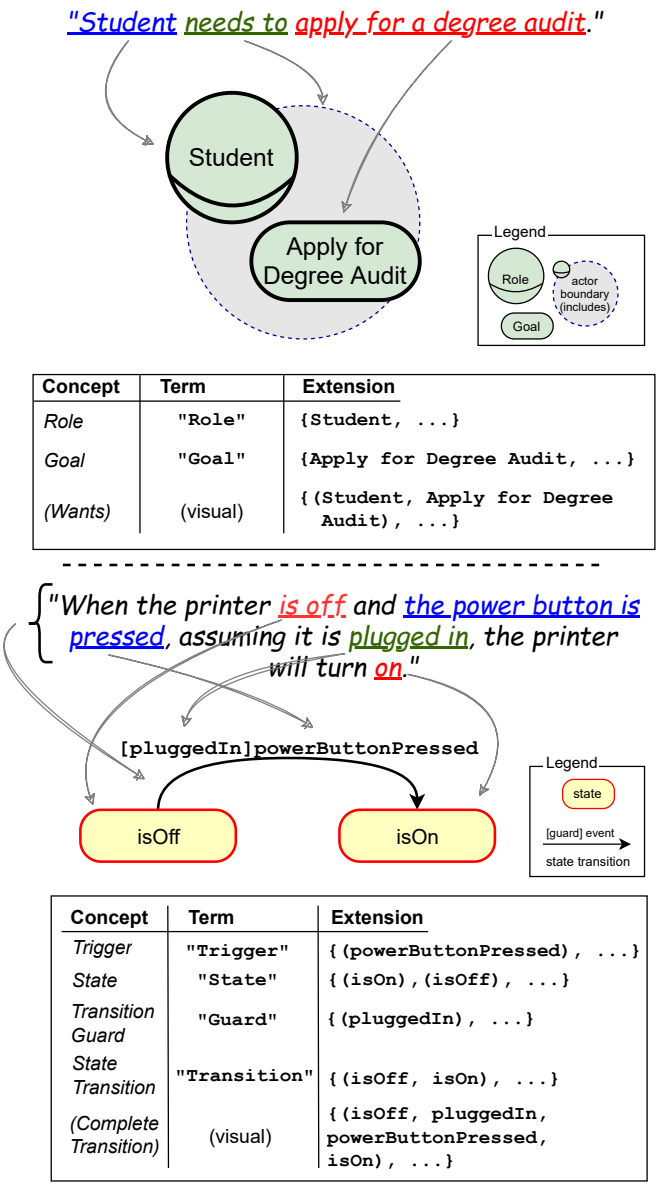

Figure 1. Model development as an extension formation process for an iStar 2.0 diagram [7] and a UML StateMachine diagram [6].

2.3 Conceptual Modeling as a Qualitative Coding Process

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, who are meant to use the language to successfully communicating among themselves.

Looking deeper into how modelers describe reality using modeling languages, 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 or enterprise 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 terms and/or visual signifiers that are available in the language.

When modeling is diagrammatic, this is done every time, for example, a new shape is placed on the diagram and text is written in it. The shape form usually signifies a concept – in turn, expressed by a term – as per the visual grammar of the language. Two such examples are depicted in Figure 1. The upper part of the diagram depicts the translation of a chunk of domain information to a model part of an diagram in iStar [7], a language for representing user intentions in requirements engineering. The lower part shows the same

process for a UML StateMachine Diagram [6]. 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 combination of a guard and a trigger.

2.4 Inter-rater agreement and conceptualization-vocabulary alignment

Language designers desire that all users of the language understand it in the same way – preferably, the way of the designers. 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 would hopefully perform the same classifications and produce the same or similar model. Reversely, by reading the model one would ideally be able to reproduce the same domain information that lead to its development. Further, the designers themselves would want to ideally agree with the way their terms are used for modeling the domain by users of the language. Detection of disagreements in the way classifications are performed, firstly among the users of the language and then between users and designers obviously constitutes a problem, as it defeats the purpose of clear communication that languages have.

Agreement or disagreement among agents with respect to how they classify subjects to categories can be quantified in ways that are well studied in the literature on qualitative content analysis [11]. In content analysis, units of content (e.g., text, images, audiovisual segments) taken from a domain are classified by raters under distinct categories or codes, so that the latter can then be used for the development of theories that are grounded on the domain information that was coded. For the purpose, various quantitative measures of *inter-rater agreement* have been proposed [12]. Absent or low observed inter-rater agreement may be due to the choice of codes and discourages further utilization of data for inferences, on the basis of lack of *reliability* of the coding (i.e., the *measurement*) process.

The coding practices employed in qualitative content analysis are analogous to modeling: in both cases raters classify domain information under terms or other signifiers representing concepts. 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. We use the term *conceptualization-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, they way it is explained and/or exemplified, or, at a deeper level, with the choice of the concept it is meant to represent.

2.5 Wand and Weber's misalignment characterization framework

One way of characterizing conceptualization-vocabulary misalignment is through the Wand and Weber framework for comparing ontological with grammatical/design constructs [13]. The authors distinguish between the set of real-world constructs we want to represent (in our case: concepts) 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 correspond to some concept.

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*, i.e. there is at least one concept in the conceptualization that cannot be 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*). That is not satisfied when a signifier in the vocabulary represents more than one concepts – so it is unclear which one 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 conceptualization-vocabulary alignment, can be useful for identifying, discussing and fixing issues with candidate language vocabularies. If language designers are told that the language suffers from, e.g., 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 construct and consider removing it from the language.

In the Section that follows we present a set of constructs for extracting misalignment indications from domain content classifications performed by samples of language users.

3 Measures of Misalignment

3.1 Overview

The measurement framework in consideration requires the following components.

- A a set of signifiers V, i.e., a vocabulary that needs to be evaluated.
- A set of human raters P who will 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, 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 development and/or evaluation uses to refer to its conceptualization. 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 that has vocabulary V. 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) are drawn from $\mathbf{D} \times E$.

As an example, consider a proposed goal-based enterprise modeling language that includes the terms *"agent"*, *"assessment"*, *"goal"*, *"has"*, *"motivates"*, *"plays-role"*, among others; the example vocabulary is inspired by iStar 2.0 [7] and Archimate [14]. These five terms constitute the vocabulary V to be evaluated. Seen as predicates, the first three are unary and the last two binary, i.e. represent relationships. 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 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. An example can be seen in Figure 2. From the description e, the evaluators have identified a set $\mathbf{D}$ of distinguished n-tuples of discourse elements. By doing so the evaluators demonstrate their belief that these elements in $\mathbf{D}$ should be modeled using terms from V. 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 are unfamiliar with 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 rater 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"*.

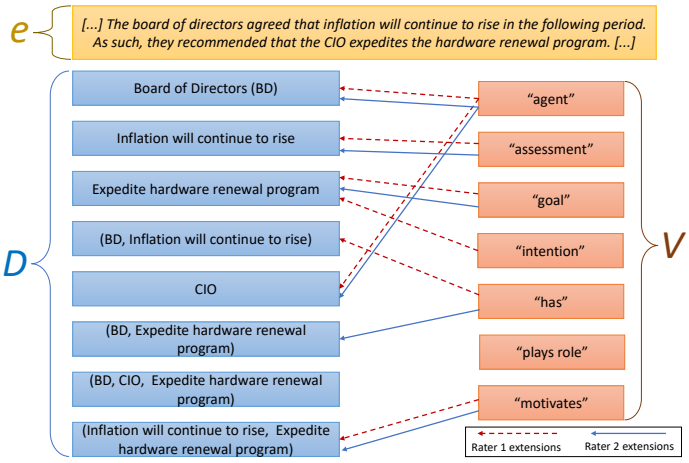

Figure 2. Rating example.

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 agree on the extension of term "has", 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, 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 classification target subjects (pairs of elements and descriptions) rather than just elements.

Observing now 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 [9], let:

- $I_p(v, e) \subset \mathbf{D}$ be the extension that rater p gave to signifier v under description e.
- $X_p(v) = \bigcup_{e \in E} I_p(v, e)$ be all the subjects that rater p classified under v.
- $B = \{s \in \mathbf{D} \times E \,|\, \exists p \in P, \exists v \in V, s.t.\ s \in X_p(v)\}$ be all subjects such that there is a rater that classified them to a signifier.
- $N \subseteq P \times \mathbf{D} \times E \times V$ be the set of all instances of a rating of a subject to a signifier by a participant.
- $N_d = \{r \in N, d \in \mathbf{D} | \, r \text{ involves } d\}$ be the set of all rating instances that involve the classification of 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:
$I_{p_2}(\text{"agent"}, e) = \{$(Board of Directors), (CIO)$\}$, $X_{p_1}(\text{"has"}) = \{$((BD, Expedite hardware renewal program),e)$\}$, $R((\text{CIO}, e), \text{"agent"}) = \{p_1, p_2\}$, $R((\text{CIO}, e), \text{"goal"}) = \{\}$, $R((\text{BD}, \text{Expedite hardware; renewal program}), e), \text{"has"}) = \{p_1\}$, and B contains all subjects except (BD, CIO Expedite hardware renewal program)

With the above functions at hand we can go ahead and define two types of abstract measurement constructs: a set for measuring *within-rater misalignment* and a set of measuring misalignment between raters and designers (*rater-authoritative misalignment*).

3.2 Within-rater Misalignment

We define the following constructs for within-rater misalignment. All constructs have two aspects whereby they indicate the issue. The *intensity* aspect refers to the depth of the issue and the *prevalence* aspect refers to how common the issue is in the data.

Construct Deficit. We have evidence of construct deficit when there is a set of one or more elements $D_{\texttt{NA}} \subset D$ such that $|N_d|$ is zero or small near zero $\forall d \in D_{\texttt{NA}}$. The larger the $D_{\texttt{NA}}$ and the smaller the N_d's the more (prevalence) and, respectively stronger (intensity) the evidence.

In our example of Figure 2, for $d =$ (BD, CIO, Expedite hardware renewal program), $|N_d| = 0$ while for $d =$ (BD, Expedite hardware renewal program) and $d =$ (BD, Inflation will continue to rise) $|N_d| = 1$. In operationalizations it is natural to also compare these values to the maximum, in our case 2.

Construct Excess. Let a signifier $v \in V$, which we wish to investigate it it is excessive. Let $B_v^{\emptyset} = \{s \in B | \, |R(s, v)| \leq c, \text{small}\, c\}$ be the set of subjects that were classified, but rarely under v. When $|B \setminus B_v^{\emptyset}| \leq k$ for some small $k \geq 0$, 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. The smaller the c and k quantities the stronger the intensity and prevalence of the issue, respectively.

In our example, there is no subject that has been classified under term "plays role", i.e., $\forall s \in B, R(s, \text{“plays role”}) = \{\}$. 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, according to the evaluators. In addition, for $c = 1$, the term "intention" would be deemed excessive: the term is used only once by p_1 to classify one subject.

Overlap. Before we discuss construct redundancy we first describe the notion of a *conceptual overlap* [9]. Consider two distinct signifiers v_1 and v_2. Consider also a subject $s \in B$. Assume that for several pairs of raters i, j, distinct or otherwise, $s \in X_{p_i}(v_1)$ and $s \in X_{p_j}(v_2)$, 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. The more the subjects and rater pairs for which this happens the more the intensity of said conceptual overlap between v_1 and v_2.

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. Let subject $s \in \mathbf{D} \times E$ 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. Let also $O(v_1, v_2)$ some measure of overlap intensity between two different signifiers v_1 and v_2. When $O(v_1, v_2)/geqc$ with respect to all or most subjects in $L_{v_1} \cup L_{v_2}$, this is an indication of construct redundancy for either of the two terms. In other words, there is no subject relevant to the two terms for which there is clarity as to which term best describes the subject. The greater the c (the minimum overlap) the greater the intensity and the more the subjects where it is observed the more the prevalence.

3.3 Rater-authoritative Misalignment

A set of constructs 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 characterized based on the distance between the user ratings and the authoritative one.

Specifically between the authoritative ratings of rater p_a and the one 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 ore more subjects which p_i does not think should be classified under v, but the designers p_a think it should. As such, the 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 ore 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 $X_{p_i}(v) \backslash X_{p_a}(v) \neq \{\}$ and $X_{p_a}(v) \backslash X_{p_i}(v) \neq \{\}$; 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 ratings are disjoint.

When inspecting any case of 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 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 term v' should not be perceived to be describing.

Back to our example, assume rater one 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 deficiency: designers think this is an intention but the rater thinks it is not. The opposite, term coarseness, is true for ((BD, Expedite hardware renewal program),e), which is the one subject in the scope excess

of "has": designers think that "has" should not evoke a meaning that would allow one to reasonably classify the specific subject under the term – for example designers may have included a specialization of the term specifically for connecting agents and goals.

The above constructs pertain to the comparison of the authoritative rating with one arbitrary rating for the purposes of demonstrating the construct's meaning. It is up to concrete operationalizations of the constructs to express the exact quantitative metrics for multiple raters.

4 Experimental Design

4.1 Overview, Research Questions and Methodology

So far we have presented a set of abstract constructs that we claim can measure, after being properly operationalized, language vocabulary design issues. To empirically test our claims we performed an experimental study with human participants. The research questions we wish to answer with our experiments are organized in two groups:

- **RQ1:** can the abstract constructs be appropriately operationalized for a data collection instrument? (**RQ1.1**) Do the operationalizations successfully detect all known issues with the languages under comparison? (**RQ1.2**) Do the operationalizations indicate issues that are known not to exist? (**RQ1.3**)
- **RQ2:** do self-reported measures of deficit, excess and overlap/redundancy correlate with observational rating-based measures in a way that one can be potentially considered as a criterion for the other?

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 intentions. 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 manufactured. To answer **RQ1** we observe whether application of the metrics will reveal the expected/planted alignment issues (**RQ1.2**) while not offering false indications of issues that were not manufactured or known to exist (**RQ1.3**). **RQ1.1** is assessed qualitatively on the basis on whether producing an operationalization was at all possible and indirectly based on whether it led to successful outcomes. Finally, **RQ2** is assessed through simple comparison of observational and attitudinal data. The purpose is both to offer further evaluation evidence vis-á-vis the metrics and to identify attitudinal questions that can potentially be used as a substitute of a full fledged classification exercises.

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

4.2 Experimental Units

The experimental units we consider in 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 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*, an existing language for modeling goals, extended with concepts *belief*, *believes-that*, *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*.

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 organics 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, inspired in part by an example found in the iStar 2.0 guide.

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) are entities and 21 (twenty-one, 9 from Heather's case and 12 from Kim's case) are binary relationships (i.e., tuples). The elements are distinct across cases, so the set of all subjects reduces to $\mathbf{D}$. The authors also define the *authoritative* way by which the elements are supposed to be classified/modeled. Their distribution can be seen in Table 1. In Table 2 the total elements per concept can be viewed for each language.

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

Table 1
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

	Goal Models		Intention Models	
	Concept	# Elements	Concept	# Elements
1	Actor	5	Organization	5
2	Goal	11	Goal	11
3	Belief	8	Objective	8
4	wants-to	7	wants-to	7
5	believes-that	4	intends-to	4
6	motivates	5	prevents	5
7	is-a-means-to	5	is-a-means-to	5

Table 2
Number of elements by language and authoritative designation.

instruments are survey-like on-line sequences of screens, and in each screen the participants are presented with information and/or ask for their input. Two such instruments are developed, one for each language. Participants are meant to be 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: 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. For simplicity, unary concepts, referred to as "entities" in this paper, are described as "concepts" in the training.

Comprehension Test: A set of questions assessing attendance and comprehension of the videos. Specifically, participants 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 already designate what element should be modeled as a unary concept and what as a binary relationship. Below each element the inventory of entities or relationships of the language in question is offered, augmented with the "`None`" 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, the two rating pages 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 rating exercises are completed, participants are asked to evaluate the language in various ways. In three (3) separate screens participants are asked in sequence 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 100 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 the concepts appearing 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 extend to which they "refer to the same thing".

Demographics: In the last screen participants offer demographic information.

For this experiment, participants are Mechanical Turk workers with a US bachelor degree and job function in Information Technology, according to self-reported data.

4.3 Operationalizations

Let us now discuss the operationalizations that are suitable for the particular instruments, by also relating the experimental material with the formalizations introduced earlier. 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))*$, \ldots\}$. 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. Hence the set of subjects is identical to the set of elements $\mathbf{D}$.

Let function $n : P \times (\mathbf{D} \times E) \times V \mapsto \{0, 1\}$, be $n(p, s, v) = 1$ if rater $p \in P$ has classified $s = (d, e)$ under v, and $n(p, s, v) = 0$ otherwise; where $v \in V$ and $s \in (\mathbf{D} \times E)$.

Denote the marginal sums:

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

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

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

Then we can operationalize the constructs we discussed earlier as described in the following subsections.

4.3.1 Construct Deficit

Recall that N_d is the set of classification instances that involve d, and that we have a construct deficit indication when for one or more d the set is empty. In our instrument in which all d's need to be rated, $|N_d| = n(\cdot, (d, \cdot), \cdot) - n(\cdot, (d, \cdot), v_{\texttt{NA}})$, recalling that $v_{\texttt{NA}}$ is a way for a rater to say she does cannot classify a presented subject to any extension.

In our operationalization we choose to compare this quantity to the total number or ratings $n(\cdot, (d, \cdot), \cdot)$ that involved d by dividing by said quantity. 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: $n(\cdot, (d, \cdot), v_{\texttt{NA}})/n(\cdot, (d, \cdot), \cdot)$. It can easily be seen that the smaller the N_d the greater the above metric as dictated by the construct's definition.

To measure the prevalence of this indication in more than one d's, we consider the array of such values for all $d \in \mathbf{D}$ and examine the maximum (if one d suffice as evidence) or a large quantile (if we want more than one d's to conclude deficit) of these values. Hence the final metric is:

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

where $Q^c(\mathbf{X})$ is the c-th percentile of the set $\mathbf{X}$, e.g., $c = 100$ (so: max) or 90 (90th percentile), and, again, $v_{\texttt{NA}}$ is a vocabulary element representing the "None" option. As per the construct stipulation, prevalence of deficit-demonstrating elements (the size of $D_{\texttt{NA}}$ in Section 3) can be detected by lowering the quantile and observing if deficit value remains high.

We also define the per participant deficit using the same ratio:

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

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

4.3.2 Construct Excess

As discussed in Section 3, construct excess is based on calculating for each $v \in V$ the set: $B^{\emptyset}_v = \{s \in B \mid |R(s,v)| \leq c, \mathrm{small}\, c\}$. Observe that in our case $|R(s,v)| = n(\cdot,s,v)$ and is bounded by $|P|$. Hence, if the quantity $U(s,v) = n(\cdot,s,v)/|P| \leq c$ for some small c for some s, this is evidence of v being excessive with respect to s, where c measures intensity. The more the subjects in $\mathbf{D} \times E$ in which $U(s,v)$ is small the more the prevalence. We again use a quantile to see if this is observed in few, most or all subjects. Hence, the comprehensive metric below:

$$\mathrm{exc}_V(v) = 1 - Q^c(\{U(v,s) \mid s \in \mathbf{D} \times E\})$$

... is close to 1 when v is excessive. Q^c is the c-th quantile as above – likely $c = 100$ in practice, or lower for more tolerance to less 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 $|\mathbf{D} \times E|$. Then:

$$\mathrm{exc}^{pp}_V(p,v) = 1 - Q^c(\{\frac{n(p,s,v)}{|\mathbf{D} \times E|} \mid s \in \mathbf{D} \times E\})$$

A statistical comparison requiring i.i.d. values would need the same set of subjects be rated by a different group of participants. In our experiment we can compare the samples:

$\mathbf{exc}^{pp}_{V_{gm}}(\text{“}motivates\text{”}) = \{\mathrm{exc}^{pp}_{V_{gm}}(p, \text{“}motivates\text{”}) \mid p \in P_{gm}\}$

$\mathbf{exc}^{pp}_{V_{im}}(\text{“}prevents\text{”}) = \{\mathrm{exc}^{pp}_{V_{im}}(p, \text{“}prevents\text{”}) \mid p \in P_{im}\}$

where *“motivates”* and *“prevents”* are as we saw (misguided) alternative terms in our two competing vocabularies.

Finally, for a per-participant measure of the excess of an entire vocabulary $v \in V$ we can simply average over the individual constituent concepts:

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

As above, in our example, samples $\mathbf{exc}^{pp}_{V_{gm}} = \{\mathrm{exc}^{pp}_{V_{gm}}(p) \mid p \in P\}$ and $\mathbf{exc}^{pp}_{V_{im}} = \{\mathrm{exc}^{pp}_{V_{im}}(p) \mid p \in P\}$ can be statistically compared for an overall indication of excess between the two vocabularies.

4.3.3 Construct Redundancy

We 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:

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

Let $L_{v_1,v_2} = L_{v_1} \cup L_{v_2} = \{s | s \in \mathbf{D} \times E \wedge ((n(\cdot,s,v_1) \geq t \times n(\cdot,s,\cdot)) \vee (n(\cdot,s,v_2) \geq t \times n(\cdot,s,\cdot)))\}, t \in [0,1]$, be the set of subjects in $\mathbf{D} \times E$ for which at least a fraction t of the total classifications they received was under v_1 or v_2; i.e. the set of subjects that are *relevant* with respect to either v_1 or v_2.

Let then $\mathbf{ov}_V(v,v') = \{ov(s,v,v') \mid s \in L_{v,v'}\}$ be the set of overlap measures between two arbitrary signifiers v and v' over all relevant subjects. An average (or other statistic, including maximum, minimum, median or quantiles) over those values offers a descriptive indicator of overlap $\mathrm{ov}_V(v,v') = mean(\mathbf{ov}_V(v,v'))$ between v and v', for, e.g., visualizations.

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 to see if there are few, many, or all of those values that are actually high.

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:

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

and likewise the overlap between v and v' according to participant p can be the average over subjects of the observed overlaps:

$$\mathrm{ov}^{pp}_V(p,v,v') = mean(\{ov(p,s,v,v') \mid s \in \mathbf{D} \times E\})$$

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 $\mathbf{ov}^{pp}_{V_{gm}}(v, v') = \{\text{ov}^{pp}_{V_{gm}}(p, v, v') | p \in P_{gm}\}$ with $\mathbf{ov}^{pp}_{V_{im}}(v, v'') = \{\text{ov}^{pp}_{V_{im}}(p, v, v'') | p \in P_{im}\}$, 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 vis-à-vis its corresponding concept. 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 $exc(\cdot,\cdot,\cdot)$. The raw numbers can be used to develop Euler diagrams for visualizing the quality and level of misalignment.

Additional indicators of the accuracy of the language that can be derived from the above are *precision* and *recall*. Recalling that precision is the number of "true positives" $acc(\cdot,\cdot,v)$ over the total number of "positives" $acc(\cdot,\cdot,v) + exc(\cdot,\cdot,v)$ and that recall is the number of "true positives" over the total number of "true" elements $acc(\cdot,\cdot,v) + def(\cdot,\cdot,v)$ it follows that:

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

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 excess of a term or of an entire vocabulary, respectively. For descriptive purposes, summary statistics can be calculated as follows:

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

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. Each of the identified hypothesis is evaluated by qualitative means 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 to the testable hypotheses of Tables 4, 7 and 10.

General indicators. In general, the manufacture intention models are expected to have an overall lower accuracy (**A.1** and **A.2**), higher deficit (**D.1.**, **D.2.**), 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 4).

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 10) 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 that "prevents" replaces "motivates". It will also present high excess (**E.2**, **E.3**).

Keeping *"is-a-means-to"* and *"wants-to"* as is. We should expect that all measures related to the concepts are similar between the two languages (**A.7**, **E.4**).

Correlations with self reported. 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**).

Of all the hypotheses (Tables 4, 7 and 10), those in the groups **A**, **D**, **E** and **O** attempt to answer **RQ1**, while hypotheses in **S** address **RQ2**.

5 PILOT STUDY[1]

A pilot study is first conducted for instrument validation. Given that the study produced did not reveal many serious issues with the instruments we here report the outcomes.

The input from 24 participants is solicited, 12 is assigned to each treatment. Three (3) and one (1) participant is excluded in each of the two treatments, goal and intention models, respectively, due to them failing three (3) or more simple comprehension questions of the videos. The gender and age characteristics of participants can be found in Table 3. All participants declare, as expected, that their educational and/or professional background is in Science, Technology or Engineering and they are native speakers of English with self-declared Excellent reading and writing skills.

Condition	Sex	Age	Count
Goal Models	Female	45	1
	Male	44	7
	Other	31	1
Intention Models	Female	46	4
	Male	41	7

Table 3
Participants Counts and Ages by Condition and Sex.

5.1 Accuracy

The results for accuracy can be found in Tables 5 and 6 for entities and relationships, respectively. In the tables the recall and precision measures for each concept are given. The 95% confidence interval presented in the brackets is created by bootstrapping 1000 times the authoritative set (all responses authoritatively designated to each concept) and the response set (all responses to the specific concept) and acquiring the corresponding quantiles.

Recall that we expect that goal model concepts will evoke more accurate responses (**A.1**) for most concepts and the overall metric (87.9% and 87.9% vs. 43.2% and 58.6%, recall and precision for the two languages goal models and intention models, respectively). We find this to be the case for most concepts and collectively the difference to be statistically significant (**A.2** - Wilcoxon $p < 0.001$ for both precision and recall on the per-participant metrics). There are two interesting exceptions: *"actor"* exhibits (somewhat) less precision than *"organization"* (i.e., the term *"actor"* is overused compared to *"organization"*). There is, further, an unexpectedly high precision value for *"goals"* in intention models: we expected that many elements authoritatively defined as *"goals"* will be in fact classified as *"objectives"*, yielding lower precision than the one observed. Further as expected *"is-a-means-to"* enjoys similar levels of accuracy across the two languages.

Pairwise comparisons between concepts also turn out as expected with one exception. While the concept *"organization"* is expected (**A.3**) to have lower recall and larger scope deficiency than the concept it replaces (*"actor"*), although the pattern vaguely appears descriptively, as seen in Figure 3, the difference in recall is not statistically significant. On the contrary, all hypothesized precision and recall differences between *"belief"* and *"believes-that"* and their replacements *"objective"*, *"intends-to"* are clearly observed and found to be statistically significant Wilcoxon $p < 0.001$ (**A.4** - **A.5**).

Lastly the *"prevents"* relationship, which has been deliberately added to trigger construct excess, is found to have low recall compared to "motivates" (**A.6**), $p < 0.005$; but precision is not low. This can best be seen in the Euler accuracy diagram of Figure 4. To explain the observed precision of "prevents", we first note that the two alternative terms *"prevents"* and *"motivates"* can be confused by a hasty and superficial participant. Thus, the few participants of the intention models that considered using the term *"prevents"*, did so in elements indeed authoritatively designated as classifiable under that term, as the term appears to fit at a first glance despite actually being semantically wrong at closer examination. For example element <*"revenue level is lower than that of the competition"*, *"try to increase sales"*>), would lead hasty raters to classify it under *"prevents"* instead of *"None"*, which is correct according to the misguided logic of the hypothetical designers of the intention models.

Finally we observe that the relationship *"is-a-means-to"* appears to be unaffected by the other changes in the language in terms of accuracy (**A.7**).

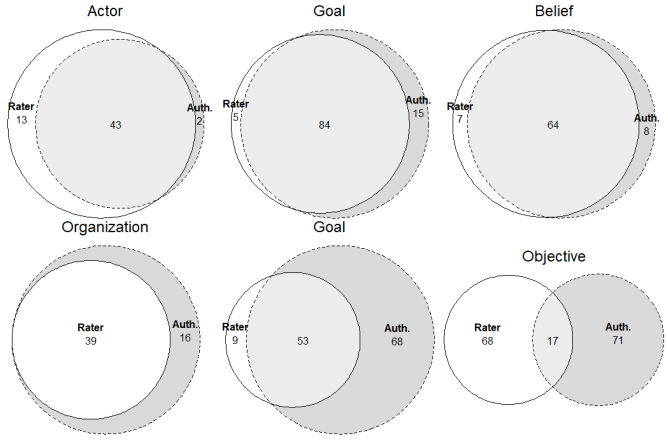

Figure 3. Euler accuracy diagrams for unary concepts

5.2 Construct Deficit

In terms of construct deficit we hypothesize that intention models will exhibit higher deficit than goal models. This is due to the absence terms *"belief"* and *"believes-that"* and

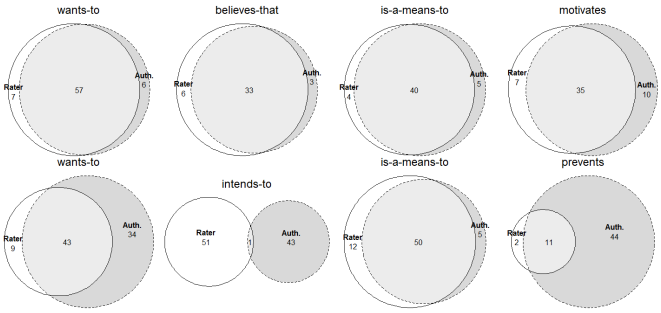

Figure 4. Euler accuracy diagrams for relationships

1. All acquired data as well as analysis scripts in R can be found in the accompanying reproducibility package [15].

	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 relationship *"is-a-means-to"*.	$\text{prec}_{V_{gm}}(v_1) > \text{prec}_{V_{im}}(v_2)$ and $\text{rec}_{V_{gm}}(v_1) > \text{rec}_{V_{im}}(v_2)$ for most $v_1 \in V_{gm}, v_2 \in V_{gm}$ except *"is-a-means-to"* $\text{prec}_{V_{gm}} > \text{prec}_{V_{im}}$ and $\text{rec}_{V_{gm}} > \text{rec}_{V_{im}}$	☆ (with comments) ☆
	A.2.	The per-participant accuracy will be higher for goal models than for intention models.	$\mathbf{prec}^{pp}_{gm} >_m \mathbf{prec}^{pp}_{im}$ and $\mathbf{rec}^{pp}_{gm} >_m \mathbf{rec}^{pp}_{im}$	☆★
	A.3.	The *"organization"* concept will have less per-participant recall and more deficiency than that of the concept it replaces.	$\mathbf{rec}^{pp}_{gm}(\text{"}actor\text{"}) >_m \mathbf{rec}^{pp}_{im}(\text{"}organization\text{"})$ *"organization"* has higher deficiency than *"actor"*	☆✪ ☆
	A.4.	Concept *"belief"* has higher per-participant precision and recall than its replacement *"objective"*.	$\mathbf{prec}_{gm}(\text{"}belief\text{"}) >_m \mathbf{prec}_{im}(\text{"}objective\text{"})$ $\mathbf{rec}_{gm}(p, \text{"}belief\text{"}) >_m \mathbf{rec}_{im}(p, \text{"}objective\text{"})$	☆★
	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"* will exhibit low recall compared to its counterpart *"motivates"*.	$\mathbf{rec}^{pp}_{gm}(\text{"}motivates\text{"}) >_m \mathbf{rec}^{pp}_{im}(\text{"}prevents\text{"})$	☆★
	A.7.	All accuracy measures of the concept *"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 4
Experimental hypotheses: accuracy.
$\mathbf{a} >_m \mathbf{b}$ is defined as $mean(\mathbf{a}) > mean(\mathbf{b})$ for samples $\mathbf{a}$ and $\mathbf{b}$
Outcomes: Observed Qualitatively (☆), Tested statistically / Passed (★),Tested statistically / Failed (✪)

	Concept	Recall	Precision
Goal Models	Actor	95.56 [91.62,99.9]	76.79 [66.14,85.09]
	Goal	84.85 [77.08,90.26]	94.38 [88.66,97.88]
	Belief	88.89 [80.74,94.63]	90.14 [81.38,94.91]
Intention Models	Organization	70.91 [60.59,79.74]	100 [100,100]
	Goal	43.8 [36.13,51.44]	85.48 [75.56,92.71]
	Objective	19.32 [11.89,28.82]	20 [12.43,29.5]

Table 5
Accuracy measures for concepts (%). In brackets the 2.5th and 97.5th percentiles of the bootstrapping distribution created by re-sampling with replacement the authoritative and response sets 1000 times and recalculating the measures.

	Concept	Recall	Precision
Goal Models	wants-to	90.48 [83.38,95.64]	89.06 [79.75,94.32]
	believes-that	91.67 [80.31,97.92]	84.62 [72.08,93.23]
	motivates	77.78 [63.82,87.8]	83.33 [69.6,91.75]
	is-a-means-to	88.89 [78.44,95.94]	90.91 [78.05,96.4]
Intention Models	wants-to	55.84 [45.73,64.85]	82.69 [71.01,91.17]
	intends-to	2.27 [-0.33,2.78]	1.92 [-0.28,2.28]
	prevents	20 [11.81,31.47]	84.62 [66.02,99.72]
	is-a-means-to	90.91 [81.68,96.66]	80.65 [70.16,87.96]

Table 6
Accuracy measures for concepts (%). In brackets the 2.5th and 97.5th percentiles of the bootstrapping distribution created by re-sampling with replacement the authoritative and response sets 1000 times and recalculating the measures.

addition of concept *"motivates"* 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 8 where entities are separated from relationships. The values in the table clearly indicate (**D.1**) the increased deficit in intention models, offering evidence that the inclusion of the "None" option serves the purpose of identifying it.

A more unexpected result is the relatively high deficit observed for goal model concepts. Drilling down we find that out of the 24 items the following four (4) have non-zero incompleteness: *"fill in an online form"* (item deficit: 0.556) *"revenue level is lower than that of the competition"* (0.333) *"use a stamp card loyalty points system"* (0.222) and *"On-line travel agency"* (0.111). There does not seem to be an obvious pattern that indicates why these particular items scored high

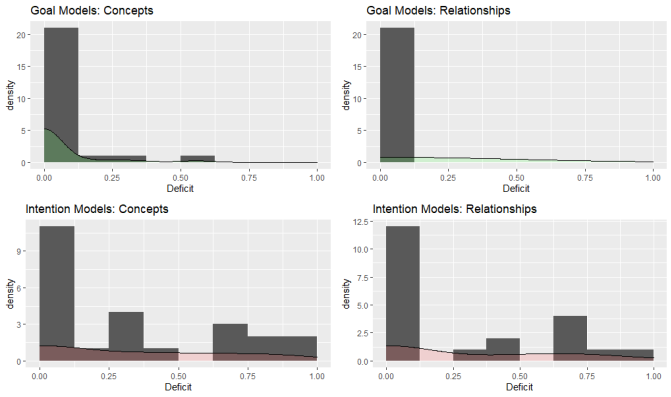

Figure 5. Distribution of item deficit values for concepts of the two languages.

in deficit.

Graphical representations such as the histogram-density plots of Figure 5, can be used to demonstrate the prevalence of deficit across items. We can, further, statistically compare the per-participant deficit between the two languages (**D.2**). Indeed a Wilcoxon rank-sum test shows that for both entities ($W = 4.5, p < 0.001$) and relationships ($W = 0, p < 0.001$) intention models exhibit less deficit, with difference between means 0.26 and 0.25, respectively.

5.3 Construct Excess

The construct excess calculated for each of the concepts of the languages can be viewed on Table 9. The indicators are as expected, with a significant difference between the two languages $W = 92, p < 0.05$ (**E.1**). In terms of individual excess measures, they highest is for *"prevents"* in intention models (0.545) and some other non-zero indications appear in concepts *"goal"*,*"objective"*, *"wants-to"* and *"intends-to"* due to the presence of overlap, which some participants consistently using mostly one of the concepts in the overlapping pair. These are 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

	H_1	Description	Tests	Outcome
Deficit	**D.1.**	Intention models will exhibit more deficit than goal models.	$\text{def}_{im} > \text{def}_{gm}$	☆ (with comments)
	D.2.	Deficit-per-participant will be greater in intention models than in goal models.	$\mathbf{def}^{pp}_{V_{im}} >_m \mathbf{def}^{pp}_{V_{gm}}$	☆★
Excess	**E.1.**	Intention models will generally exhibit more construct excess than goal models.	$\text{exc}_{im}(v_1) > \text{exc}_{gm}(v_2)$ for various $v_1 \in V_{im}, v_2 \in V_{gm}$	☆ (with comments)
			$\mathbf{exc}^{pp}_{im} >_m \mathbf{exc}^{pp}_{gm}$	★
	E.2.	The relationship *"prevents"* clearly stands-out as excessive and so may *"goal"*, *"objective"*, *"wants-to"* and *"intends-to"*.	$\text{exc}_{im}(\text{"prevents"}) > \text{exc}_{[\cdot]}(v)$ for $v \in V_{gm} \cup V_{im}$	☆
	E.3.	The relationship *"prevents"* is more excessive than *"motivates"* which it replaces in terms of per-participant excess.	$\mathbf{exc}^{pp}_{im}(\text{"prevents"}) >_m \mathbf{exc}^{pp}_{gm}(\text{"motivates"})$	☆★
	E.4.	No strong evidence of excess emerges in unproblematic concepts, including all those of goal models and *"organization"* and *"is-a-means-to"* in intention models.	Low $\mathbf{exc}_{[\cdot]}(v)$ for the mentioned v	☆

Table 7
Experimental hypotheses: excess and deficit.
$\mathbf{a} >_m \mathbf{b}$ is defined as $mean(\mathbf{a}) > mean(\mathbf{b})$ for samples $\mathbf{a}$ and $\mathbf{b}$
Outcomes: Observed Qualitatively (☆), Tested statistically / Passed (★), Tested statistically / Failed (✪)

	Concept Type	max	q95	q90	q75
Goal Models	Entities	0.56	0.32	0.19	0.00
	Relationships	0.00	0.00	0.00	0.00
Intention Models	Entities	0.91	0.90	0.82	0.64
	Relationships	1.00	0.82	0.73	0.64

Table 8
Deficit measures: maximum observed, and 95th, 90th and 75th quantile.

a significant result – Wilcoxon $W = 95, p < 0.05$), but so do several other comparisons, primarily due to other problems (e.g., overlaps, deficits and accuracy problems).

We finally observe that the metric does not seem to be picking up false positives. All excess measures of goal models are zero and so are intention model concepts "organization" and "is-a-means-to", exactly as expected (**E.4**).

Goal Models		Intention Models		per part.
Concept	Excess	Concept	Excess	diff. p value
Actor	0.000	Organization	0.000	0.003 *
Goal	0.000	Goal	0.182	0.938
Belief	0.000	Objective	0.182	0.000 *
wants-to	0.000	wants-to	0.182	0.001 *
believes-that	0.000	intends-to	0.091	1.000
motivates	**0.000**	**prevents**	**0.545**	**0.000 ***
is-a-means-to	0.000	is-a-means-to	0.000	0.174

Table 9
Excess per concept. (*) = significant at $p = 0.05/7 = 0.007$ after Bonferroni correction for the 7 comparisons.

5.4 Overlaps and Redundancy

To explore redundancy the first exploratory step is to calculate overlaps for all pairs of concepts. For the analysis, we set a relevance threshold to $t = 0.1$, 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.

The overlap charts of Figure 6 shows average overlaps over all relevant items. All overlaps for goal models appear to be bounded by 0.16, while in intention models the overlaps between *"goal"* and *"objective"* and *"wants-to"* and *"intends-to"* stand out as exactly expected (**O.1**)

Calculations of overlaps per participant are also pertinent in this analysis. The statistical tests comparing the two benchmark pairs (**O.2**) appear to fail. Looking closely into that data, we observe that participants appear to not utilize the check-all-that-apply form of the questions, overwhelmingly picking only one of the available choices. Thus, with per-participant overlap sparse in the data, the corresponding per-participant statistical results may be unreliable.

Finally, as per the operationalizations, to specifically study redundancy of a concept we need to collect a low-quantile (over subjects) of the overlaps that the concept exhibits in relation to each other concept. Of these pairwise calculations between concepts, we are interested in all non-zero ones and particularly the maximum. Results for intention models can be found in Table 11. Thus, based on 10th percentiles, for *"goal"* and *"objective"* the corresponding redundancy indices are 0.11 and for the *"wants-to"* and *"intends-to"* 0.05. As more small values are ignored (as e.g. coincidental) by moving to higher percentiles, the more the index increases. For goal models the corresponding values are all zero. This seems to confirm our expectation **O.3**, noting, however, that, for a complete picture, practical application seems to require inspection of various quantiles as in Table 11 rather that simply calculating, e.g. the minimum of overlap values.

5.5 Self-reported Data

Let us now turn our focus to the self reported data and their correlation (or lack thereof) to the observed data.

Deficit. Participants are asked to rate the degree to which they found the concept set to be *complete* (scale [0,100]). The higher the grade the more complete they found the concept set to be. Naturally we expect a negative correlation between the observed deficit and the reported completeness (**S.1**). Indeed, for intention models, a negative (Kendall's tau) correlation is calculated for both entities (-0.353) and relationships (-0.628) but statistically significant only in the latter case $p < 0.05$ and after an outlier has been removed after graphical inspection - milder statistically insignificant negative correlations are observed with the outlier. For goal models, the analysis for relationships is not meaningful as no deficit whatsoever is observed there. For concepts, a negligible non-significant correlation is observed (0.08). The negative relationship for all data of the intention model can also be seen in Figure 7.

	$\mathbf{H}_1$	Description	Tests	Outcome
Overlap	**O.1.**	The overlaps between *"goal"* and *"objective"* and between *"wants-to"* and *"intends-to"* are substantially higher than all other pairs.	ov_{im}(*"goal"*, *"objective"*) and $\mathrm{ov}_{V_{im}}$(*"wants-to"*, *"intends-to"*) are higher than all other overlaps	☆
	O.2.	The above pairs overlap higher in intention models than in goal models after replacing the original concepts on a per-participant basis.	$\mathbf{ov}^{pp}_{V_{im}}(p,$ *"goal"*, *"objective"*$) >_m$ $\mathbf{ov}^{pp}_{V_{gm}}(p,$ *"goal"*, *"belief"*$)$ ---- $\mathbf{ov}^{pp}_{V_{im}}(p,$ *"wants-to"*, *"intends-to"*$) >_m$ $\mathbf{ov}^{pp}_{V_{gm}}(p,$ *"wants-to"*, *"believes-that"*$)$	✗✪
	O.3.	Redundancy will be 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.**	In each language, there is a negative correlation between the self reported measure of completeness and the per-participant observational measure of deficit $\mathbf{def}^{pp}_V$	...	☆✪
	S.2.	For each concept in each language, there is a negative correlation between the self reported measure of relevance and the observational measure of per-participant excess $\mathbf{exc}^{pp}_V(v)$	...	✗✪
	S.3.	There is a positive correlation between the self reported and the per-participant observational measures of overlap $\mathbf{ov}^{pp}_V(v, v')$	...	✗

Table 10
The experimental hypotheses: overlap, redundancy and relationship to self-reported data
$\mathbf{a} >_m \mathbf{b}$ is defined as $mean(\mathbf{a}) > mean(\mathbf{b})$ for samples $\mathbf{a}$ and $\mathbf{b}$
Outcomes: no evidence or low quality data (✗), observed qualitatively (☆), tested statistically / passed (★), tested statistically / failed (✪)

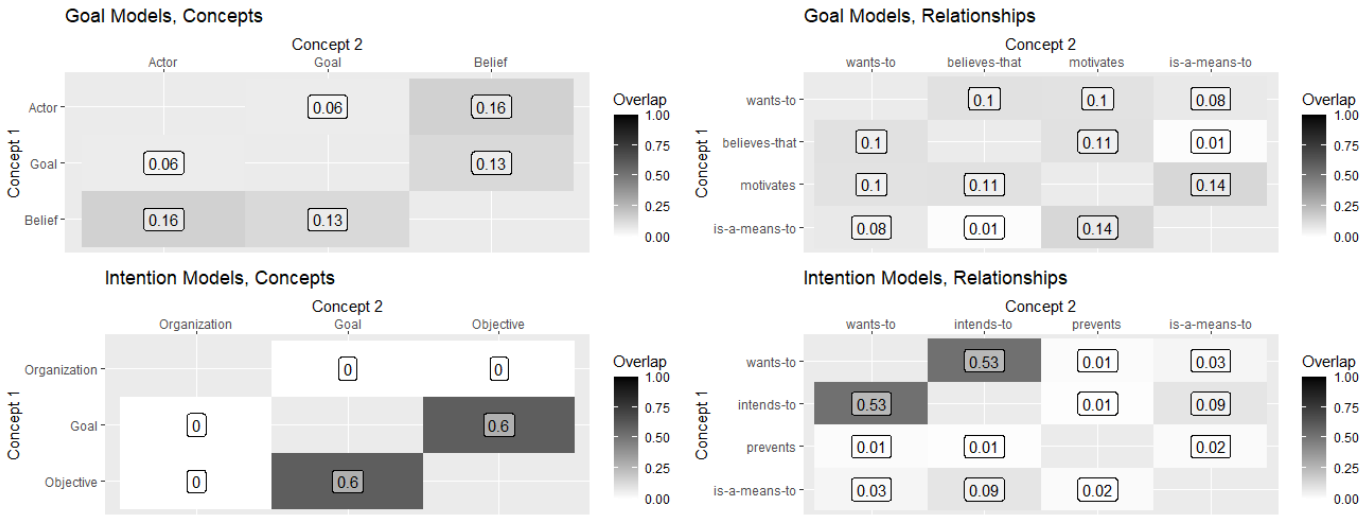

Figure 6. Overlap Chart (t = 0.1)

Concept1	Concept2	median	q25	q10	min
Goal	**Objective**	**0.70**	**0.24**	**0.11**	0.00
Organization	Goal	0.00	0.00	0.00	0.00
Organization	Objective	0.00	0.00	0.00	0.00
intends-to	is-a-means-to	0.00	0.00	0.00	0.00
intends-to	prevents	0.00	0.00	0.00	0.00
prevents	is-a-means-to	0.00	0.00	0.00	0.00
wants-to	**intends-to**	**0.62**	**0.19**	**0.05**	0.00
wants-to	is-a-means-to	0.00	0.00	0.00	0.00
wants-to	prevents	0.00	0.00	0.00	0.00

Table 11
Redundancy indexes for Intention Models (for goal models all values are zero)

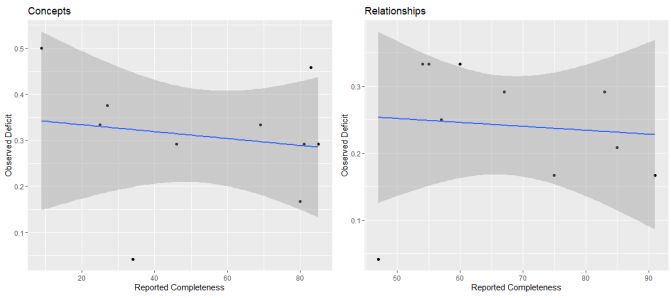

Figure 7. Observed versus self-reported assessments of deficit (intention models)

Excess. Participants are asked to rate each concept of the respective language 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**). Such correlations are not observed with any consistency, as seen in Table 7. Even for a concept such as *"prevents"* which is clearly excessive, the correlation appears to be positive; though not statistically significant. Based on looking at the data there appears to be a misconception with the question of the rating instrument, with some participants interpreting the question reversely. The particular finding warrants further investigation and alternative ways

to elicit perception of concept relevance.

Concept	τ	p	Concept	τ	p
Organization	0.44	0.17	Actor	0.12	0.77
Goal	-0.07	0.84	Goal	0.43	0.25
Objective	0.12	0.72	Belief	-0.06	0.88
wants-to	-0.36	0.28	wants-to	0.42	0.26
intends-to	-0.59	0.06	believes-that	-0.12	0.75
prevents	0.21	0.54	motivates	0.02	0.97
is-a-means-to	0.53	0.09	is-a-means-to	-0.39	0.30

Table 12
Correlations (Kendall tau τ and associated statistical significance p) between self-reported measures of relevance and observed measures of excess.

Overlap. As a last self-reporting activity participants are asked to rate the overlap between pairs of entities and relationships. We again expect these ratings to correlate positively with the measured overlaps. A visual representation of the result can be viewed in Figure 8. Notice also in the figures that, for intention models, the two manufactured overlaps are observed in both the observational and the self-reported data. An interesting feature of these results is that participants report a small overlap rating between 0.2 and 0.5 (in a scale 0..1.0 scale) even if none is observed.

Despite the qualitative evidence, the fact the participants did not utilize multiple ratings per element to specify "intra-participant" overlap a statistical analysis on a per-participant basis is not possible. Thus we remain inconclusive about **S.3**.

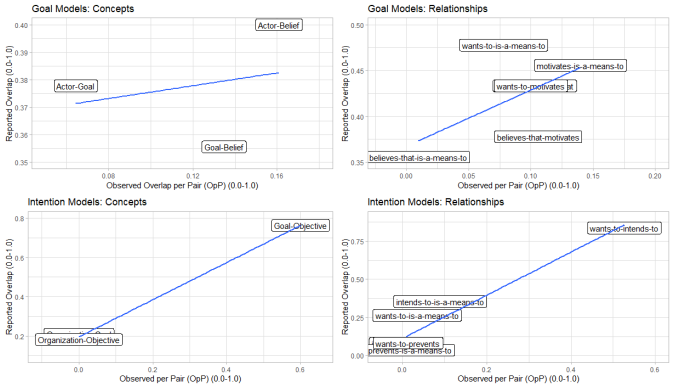

Figure 8. Relationships between self-reported and observed measures of pair-wise overlap.

6 Concluding Remarks and Next Steps

We presented EmVoc a framework for empirically evaluating modeling language conceptualization and vocabulary qualities, through observing how samples from the intended language user population use the concepts and their representations to model relevant domain phenomena. The framework is based on the definition of a set of abstract empirical constructs, which, after proper operationalization, can be used for analyzing observations in order to identify different categories of issues with the choice of concepts and/or their representations.

In this paper we presented an empirical study conducted as an initial evaluation of our framework: two languages with known issues are compared using our framework, and we test (a) whether EmVoc will offer the expected indications, (b) whether EmVoc's indications agree with how participants themselves evaluate the design choices of the designers.

The results are largely consistent with our hypotheses, with some exceptions that may be due to sub-optimal instrument design. Specifically, the following steps shall be taken in the next round of data collection:

- Participants have not utilized the ability to select multiple items from the presented inventories. As a result, per-participant overlap measures were difficult to analyze. The ability to make multiple selections must be advertized better in the instrument.
- The self-reporting instruments appear to not have been comprehended by participants. Explain and/or re-word them better.

References

[1] P. Fettke, "How Conceptual Modeling Is Used," in *Communications of the Association for Information Systems*, vol. 25, 2009. [Online]. Available: https://doi.org/10.17705/1CAIS.02543

[2] I. Davies, P. Green, M. Rosemann, M. Indulska, and S. Gallo, "How do practitioners use conceptual modeling in practice?" *Data & Knowledge Engineering*, vol. 58, no. 3, pp. 358–380, 2006. [Online]. Available: https://doi.org/10.1016/j.datak.2005.07.007

[3] A. Olivé, *Conceptual Modeling of Information Systems*. Springer, 2007.

[4] P. G. Alexander T. Borgida Vinay K. Chaudhri and E. S. Yu, Eds., *Conceptual Modeling: Foundations and Applications*. Springer, 2009.

[5] J. Mylopoulos, "Conceptual Modelling and Telos," in *Conceptual Modeling, Databases and CASE: An Integrated View of Information Systems Development*, P. Loucopoulos and R. Zicari, Eds., 1991.

[6] Object Management Group, "Unified Modeling Language (v2.5.1)," Tech. Rep., 2017.

[7] F. Dalpiaz, X. Franch, and J. Horkoff, "iStar 2.0 Language Guide," *The Computing Research Repository (CoRR)*, vol. abs/1605.0, 2016. [Online]. Available: http://arxiv.org/abs/1605.07767

[8] R. L. Rosnow and R. Rosenthal, *Beginning Behavioral Research: A Conceptual Primer*, 6th ed. NJ, USA: Pearson Prentice Hall, 2008.

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

[10] N. Guarino, D. Oberle, and S. Staab, *What Is an Ontology?*, 2009, pp. 1–17. [Online]. Available: https://doi.org/10.1007/978-3-540-92673-3_0

[11] K. Krippendorff, *Content Analysis: An Introduction to it Methodology*. SAGE, 2004.

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

[13] Y. Wand and R. Weber, "On the ontological expressiveness of information systems analysis and design grammars," *Information Systems Journal*, vol. 3, no. 4, pp. 217–237, 1993. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2575.1993.tb00127.x

[14] The Open Group, "ArchiMate® 3.1 Specification," Tech. Rep., 2019.

[15] S. Liaskos, "Replication Data for: The EmVoc framework for empirically evaluating modeling language vocabulary qualities: a pilot evaluation study," 2022. [Online]. Available: https://doi.org/10.5683/SP3/OGH0UO