Empirically Evaluating the Semantic Qualities of Language Vocabularies

Sotirios Liaskos[1] ✉, John Mylopoulos[2], and Shakil M. Khan[3]

[1] School of Information Technology, York University, liaskos@yorku.ca
[2] Dept. of Computer Science, University of Toronto, jm@cs.toronto.edu
[3] Dept. of Computer Science, University of Regina, shakil.khan@uregina.ca

Abstract. Developing and representing conceptualizations is a critical element of conceptual modeling language design. Designers choose a set of suitable concepts for describing a domain and make them the core of the language using suggestive terms that convey their meaning to language users. Additional documentation and training material, such as examples and guides, aims at ensuring that the chosen terms indeed evoke the concepts designers intended. However, there is no guarantee that language designers and users will eventually understand the correspondence between terms and concepts in the same way. This paper proposes a framework for empirically evaluating the vocabulary appropriateness of modeling languages and characterizing its absence in terms of established language design issues. The framework is based on the definition of a set of abstract empirical constructs that can be operationalized into different concrete measures, depending on study requirements and experimental design choices. We offer examples of such measures and demonstrate how they inform language design through a hypothetical language design scenario using a mix of realistic and simulated data.

Keywords: Conceptual Modelling · Conceptualization Quality · Empirical Conceptual Modelling · Goal Models

1 Introduction

Designing and representing conceptualizations lies at the very core of conceptual modeling language development. Conceptualizations are sets of concepts selected by designers as suitable for modeling a domain [6,7]. Each concept in a conceptualization is meant to capture intuitively some facet of the domain and is conveyed to users through a term that is familiar to domain users. Guides with definitions and examples are available to further ensure that the meaning of each term is shared between designers and users. However, whether such a vocabulary of terms is properly understood by users is not straightforward to determine. The chosen terms may evoke among users a meaning that is different from the one the designers designated it for, or users may be found to be confused or disagreeing among themselves about the meanings of those terms.

For example, consider $i*$, a requirements modeling language, proposed in 1995 by E. Yu [23]. The language was intended to model stakeholders and their goals, as well as ternary social dependence relationships among them. Towards this end, the language offered concepts signified through terms such as *"actor"*, *"agent"*, *"position"* and *"role"*. Since its inception, the language has been used extensively for research and teaching purposes by many research groups, largely organized around the iStar workshop series. In 2015, that community decided to conduct an evaluation of the *i** experience, and on the basis of its findings proposed iStar 2.0 [4]. One of the findings was that users, especially students, of *i** were confusing the notions of *"position"* and *"role"* in their models. As a result, *"position"* was dropped from the new concept set and vocabulary. While the *i** community had the benefit of many years of experience to inform such updates, one wonders if there is a quicker and more systematic way to empirically assess the success of a vocabulary selection for any language.

In this paper, we propose a framework for empirically measuring the vocabulary appropriateness of conceptual modeling languages. The framework is based on offering domain descriptions to representative language users and inviting them to categorize domain elements relative to the terms in the vocabulary. The framework includes a set of abstract empirical constructs for analyzing the resulting data, informed by an established model of vocabulary pathologies [7,25]. The constructs can be operationalized into concrete measures, based on the format of the data collection instruments and the needs of the study at hand. An application of the framework on a mix of realistic data from a past experiment and simulated data demonstrates how the constructs can be translated to concrete metrics and how they can indicate the type of corrections needed in the design. The work generalizes and systematizes our earlier work [15] so that it is compatible with established approaches for understanding and analyzing language qualities. It further introduces measures for additional quality issues such as construct deficit, construct excess, and construct redundancy. Moreover, thanks to a new formulation, our framework accounts for conceptual relations of arbitrary arity, rather than just unary arity [15].

The rest of the paper is organized as follows. In Section 2 we offer the necessary background on conceptualizations, in Section 3 we introduce our empirical framework and in Section 4 we describe an application thereof. We review related work and conclude in Sections 5 and 6.

2 Research Baseline

2.1 Conceptualizations, Languages and Ontological Commitments

Following [6] consider a system S that we are interested in modeling. We first define conceptual relations (aka concepts) and conceptualizations over a domain D of distinguished elements of S, given a set of possible worlds W (states of S).

Definition 1. *A conceptual relation (henceforth:* ***concept****) is a total function* $\rho^n : W \mapsto 2^{D^n}$ *from worlds to all possible extensional n-ary relations on* D*. A* ***conceptualization*** *is, then, a triple* $\mathbf{C} = (D, W, \Re)$ *in which* $\Re$ *is a set of such concepts on the domain space* $< D, W >$*.*

Such concepts need to somehow be built into a language that users can use to build models of the domain. Towards this end, language designers select an appropriate *term* (name or expression) for each concept based on its intended meaning and the cultural context where the language is meant to be used. In the goal modeling domain, for example, the concept *goal* can be represented in English using terms such as *"goal"*, *"intention"* or *"objective"*. A language is thus grounded on terms for representing concepts, matching such semantic preconceptions. Thus [6]:

Definition 2. *Let* $\mathbf{L}$ *be a language with vocabulary* V*. A model for* $\mathbf{L}$ *is a tuple* $M = (S, I)$ *where* $I : V \mapsto D \cup \mathbf{R}$*, the* ***interpretation****, maps names and terms from* V *to elements in either* D *or* $\mathbf{R}$*, the latter being a set of n-tuples from* D*. The exact subset of* $D \cup \mathbf{R}$ *to which a vocabulary element* $v \in V$ *maps is called the* ***extension*** *of* v*.*

Definition 3. *An* ***ontological commitment*** *for* $\mathbf{L}$ *is a tuple* $\mathbf{K} = (\mathbf{C}, \mathfrak{I})$*, where* $\mathfrak{I}$ *is a total function* $\mathfrak{I} : V \mapsto D \cup \Re$*, i.e., where every symbol in* V *maps to either an element of* D *or a concept in* $\Re$*. We, further, denote as* V_D *the portion of the vocabulary reserved for mapping to elements of* D *and* $V_\Re$ *the terms reserved for mapping to concepts. We will henceforth refer to elements in* $V_\Re$ *as* ***concept terms*** *(or simply* ***terms****).*

For example, for a vocabulary $V = V_D \cup V_\Re$ with $V_D = \{$ *"Alice"*,*"pay the bills"*$\}$ and $V_\Re = \{$ *"actor"*, *"goal"*, *"wants"*$\}$, an ontological commitment maps the term *"actor"* to, say, the concept of an individual who can act in a domain, the term *"goal"* to the concept of a desired state of affairs, and the term *"wants"* to the concept that an actor is inclined to pursue a goal. Models of the language can be compatible or incompatible with the commitment. For example, including *"Alice"* in an extension of term *"actor"* in a model M is consistent with the commitment. However, including *"pay the bills"* is inconsistent, as the latter does not satisfy the definition of an actor as per the ontological commitment.

To accomplish clear communication, the language designers must choose a vocabulary V that intuitively conveys the right commitment $\mathbf{K}$ to modelers and model readers. However, achieving such a shared understanding of the commitment is neither guaranteed nor trivial to assess.

2.2 Language Qualities and their Measurement

To characterize inadequate sharedness of the ontological commitment of a vocabulary we adopt a framework for language quality due to Wand and Weber [25]. The key concern in that framework is the degree of alignment between language terms and concepts, whose absence the authors characterize using four (4) different quality issues. Firstly, when there are concepts in the domain that are not represented in the vocabulary we have *construct deficit.* Secondly, if there are vocabulary terms that represent multiple concepts, we have *construct overload.* Thirdly, when there are multiple vocabulary terms that represent the same concept, this is a case of *construct redundancy.* Finally, when there are vocabulary terms that do not relate to any concept we have *construct excess.*

Language and empirical set-up:

$V_{\Re}$ = { *"goal"*, *"objective"*, *"argument"*, *"wants"*, *"desires"*}
E = {e_1 = *"Alice plans to pay her bills."*,
e_2 = *"Alice would like to pay her bills but it is not her priority now."*}
V_D = { *"Alice"*, *"pay bills"*}
D = { *"Alice"*, *"pay bills"*, ⟨ *"Alice"*, *"pay bills"*⟩}

Extensions:

	Rater:	p_1		p_2	
	Descr.:	e_1	e_2	e_1	e_2
Term	*"goal"*	*"pay bills"*	*"pay bills"*	–	–
	"objective"	–	–	*"pay bills"*	*"pay bills"*
	"argument"	–	–	–	–
	"wants"	⟨ *"Alice"*, *"pay bills"*⟩	–	⟨ *"Alice"*, *"pay bills"*⟩	⟨ *"Alice"*, *"pay bills"*⟩
	"desires"	⟨ *"Alice"*, *"pay bills"*⟩	⟨ *"Alice"*, *"pay bills"*⟩	⟨ *"Alice"*, *"pay bills"*⟩	–

Table 1. Running Example.

Let us explore how the above vocabulary issues can be empirically detected. The proposed measurement process is inspired by processes for measuring *reliability* in the context of qualitative content analysis [12,15]. In content analysis, *units* of content (text, audiovisual segments) representing information about the domain are classified by *raters* into a set of categories (*codes*) that best describe each unit. The exercise is meant to allow the development of theories about the content grounded on codes and is predicated on the presence of agreement among raters on what codes best describe each unit. Lack of inter-rater agreement implies an unreliable coding process, which can be due to a variety of factors, including problems with the appropriateness of the coding language.

To apply these ideas to our measurement problem, we have (a) samples of language users play the role of raters, (b) descriptions of elements from the domain play the role of content units, and (c) the terms in the vocabulary $V_{\Re}$ play the role of codes. As in content analysis, we ask language users to assign domain elements to one or more vocabulary terms – if applicable. Ideally, they will all agree with their assignments indicating good sharing of the ontological commitment. If not, however, the different ways by which raters disagree are indicative of different categories of issues with the choice of vocabulary, in accordance with Wand and Weber's framework. We offer more details below.

3 Empirically Measuring Semantic Qualities

3.1 Method Overview and Notation

Let us now describe more concretely the method for acquiring and analyzing vocabulary quality data with reference to the example of Table 1:

1. Identify the Language. Consider a language **L** with an ontological commitment **K** and a set of terms for representing concepts $V_{\Re} = \{r_1, r_2, \ldots\} \subseteq V$. Let $V_{\Re}$ = { *"goal"*, *"objective"*, *"argument"*, *"wants"*, *"desires"*} of the upper part of Table 1 be the vocabulary of interest for our running example.

2. Sample Raters. Select a set $p \in P$ of human raters. Selected raters are representative users of the vocabulary so to allow generalization of findings to all intended users of the language. They should also have good knowledge of the domain to ensure that their categorizations reflect the features of the language, rather than their own understanding of the domain.

3. Construct Descriptions. Construct a set of descriptions $E = \{e_1, e_2, e_3, \ldots\}$ each partially describing in natural language a world in W. Descriptions present domain phenomena that the language is meant to model. For our example, two such descriptions can be seen in Table 1, though descriptions are meant to be much more extensive in practice – see [16]. Sampling of such descriptions is biased towards descriptions that test all expressive capabilities of the language and are expected to trigger utilization of all language terms.

4. Identify Discourse Elements. Extract from the descriptions a set of discourse element representations V_D and n-tuples from that set that, according to the designers, are relevant to the domain. For our example of Table 1, we identify two elements (*"Alice"*, *"pay bills"*) and one tuple therewith ($\langle$*"Alice"*,*"pay bills"*$\rangle$). Let $\mathbf{D} \subset 2^{(V_D)^n}$ be the union of V_D with the the set of all n-tuples constructed from it.

5. Raters Form Term Extensions. For each description $e \in E$, ask each rater $p \in P$ to form the extension of each concept term $r \in V_{\Re}$ using elements from $\mathbf{D}$. As described above, the rater goes over the samples $\mathbf{d} \in \mathbf{D}$ and, for each, she decides whether it should be included in the extension of r based on the evoked concept. If yes, we say then that the rater p *classifies* $\mathbf{d}$ under r. We, further call the pair $(\mathbf{d}, e)$, i.e. an element or n-tuple of elements from V_D in a context of a description e, *subject.*

The result of a rating exercise can be seen in the lower part of Table 1. We consider two raters $p_1, p_2 \in P$. Each cell in the table describes the extension that each rater constructed for each term under each description. For example, for both e_1 and e_2, out of all elements in V_D rater p_1 classifies only *"pay bills"* under term *"goal"*. Thus, both subjects (*"pay bills"*, e_1) and (*"pay bills"*, e_2) are classified under *"goal"*. However, only subject ($\langle$*"Alice"*,*"pay bills"*$\rangle$, e_1) is in the extension of term *"wants"* according to p_1.

6. Analyze Extensions. Extensions developed in the previous step are compared and analyzed to identify and characterize problems with the proposed vocabulary. We define the constructs for such characterizations below.

3.2 Rater-based Measures of Completeness and Clarity

Denote $I_p(r, e) \subseteq \mathbf{D} \times E$ to be the extension of concept term r finally constructed by rater p given description e in Step #5 above. Consider also the union $X_p(r) = \bigcup_{e \in E} I_p(r, e)$ of all subjects that rater p classified under r. For instance, in our running example: X_{p_1}(*"goal"*) $= \{($*"pay bills"*, $e_1)$, $($*"pay bills"*, $e_2)\}$ and X_{p_2}(*"desires"*) $= \{(\langle$*"Alice"*,*"pay the bills"*$\rangle$, $e_1)\}$. Further, let $B = \{s \in \mathbf{D} \times E \mid \exists p \in P, \exists r \in V_{\Re} \text{ s.t. } s \in X_p(r)\}$ all subjects rated and $R_s(r) = \{p \in P \mid s \in X_p(r)\}$ be the subset of raters that classified s under r. For instance $R_{(\text{“}pay\ bills\text{”}, e_2)}(\text{“}goal\text{”}) = \{p_1\}$ and $R_{(<\text{“}Alice\text{”}, \text{“}pay\ bills\text{”}>, e_1)}(\text{“}desires\text{”}) = \{p_1, p_2\}$, $R_{(\text{“}Alice\text{”}, e_1)}(\text{“}argument\text{”}) = \{\}$. We then define the following constructs:

Construct Deficit: Let $B_{\tilde{d}} \subseteq \mathbf{D} \times E$ be the set of subjects that involve $\tilde{d} \in V_D$.

The greater the difference $B_{\tilde{d}} \setminus B$ the more the evidence of *construct deficit*, i.e., there are elements $\tilde{d}$ of the domain of discourse that are consistently excluded from extensions.

In our example, consider $\tilde{d} = \textit{“Alice”}$. $B_{\tilde{d}} = \{(\textit{“Alice”},e_1),\ (\textit{“Alice”},e_2),\ (\langle \textit{“Alice”},\ \textit{“pay bills”}\rangle,e_1),(\langle \textit{“Alice”},\textit{“pay bills”}\rangle,e_2)\}$ and, thus, $B_{\tilde{d}} \setminus B = \{(\textit{“Alice”},e_1),\ (\textit{“Alice”},e_2)\}$. That is, unary element *“Alice”* was not classified under any term by any rater, yet was included in V_D as an element that needs to be modeled. Thus, a new term, such as *“actor”*, may need to be introduced in $V_{\Re}$ to describe elements like *“Alice”*.

Construct Excess: Let r be one of the concept terms. If $\forall s \in B,\ |R_s(r)| \leq c$, for some small c, this is evidence of *construct excess*, with r being the excessive vocabulary construct. The smaller the c the stronger the evidence.

In our example, $R_s(\text{“}\textit{argument}\text{”}) = \{\}$ for all $s \in B$, as in neither of the descriptions has any of the raters classified any element of $\mathbf{D}$ under *“argument”*. This is a symptom of *“argument”* being an excessive term, i.e. a term that is not useful for representing something of interest in the domain.

Overlaps: Let subject $s \in \mathbf{D} \times E$ and two terms r_i and r_j such that $r_i \neq r_j$. Assume that for several pairs of raters l, m (possibly the same rater $l = m$), $s \in X_{p_l}(r_i)$ while $s \in X_{p_m}(r_j)$ – so s is classified both under r_i and under r_j by the same or by different raters. We say that this is a *conceptual overlap* between r_i and r_j with respect to s. The more the instances of such classification divergence between r_i and r_j over s, the more the overlap between r_i and r_j over s.

In our example, there is no subject that is classified both under *“goal”* and under *“wants”* by the same or different rater. Hence, there is no overlap between those terms. However, between *“wants”* and *“desires”* there is an overlap with respect to $(\langle \textit{“Alice”},\ \textit{“pay bills”}\rangle,\ e_2)$ and $(\langle \textit{“Alice”},\ \textit{“pay bills”}\rangle,\ e_1)$, as both subjects are assigned in the extensions of both two terms.

Construct Redundancy: Let a subject $s \in \mathbf{D} \times E$ be *relevant* to a construct r, if a minimum number of raters have included the subject in the extension of r. Assume that two different constructs r_i, r_j almost always overlap with respect to any subject that is relevant to either of them. This is an indication of *construct redundancy* of r_i or r_j, i.e., according to the raters, whenever any of the two terms is used, the other term could have been used as well.

In our example, $(\textit{“pay bills”},e_1)$ and $(\textit{“pay bills”},e_2)$ are the only subjects that are relevant to *“goal”* and *“objective”*. The two terms overlap with respect to both subjects due to inter-rater disagreements. There is no subject relevant to the terms with respect to which there is no overlap. Thus, we can mark either term *“goal”* and *“objective”* as possibly redundant.

In Figure 1, left side, an abstract schematic representing the logic of the above is shown. The inclusion of dots to frames represents the frequent inclusion of subjects to the corresponding term extensions.

3.3 Measures in the Presence of Authoritative Ratings

Consider now that one of the raters p_a is the designer of the language, i.e., the agent that develops the vocabulary on the basis of the ontological commitment

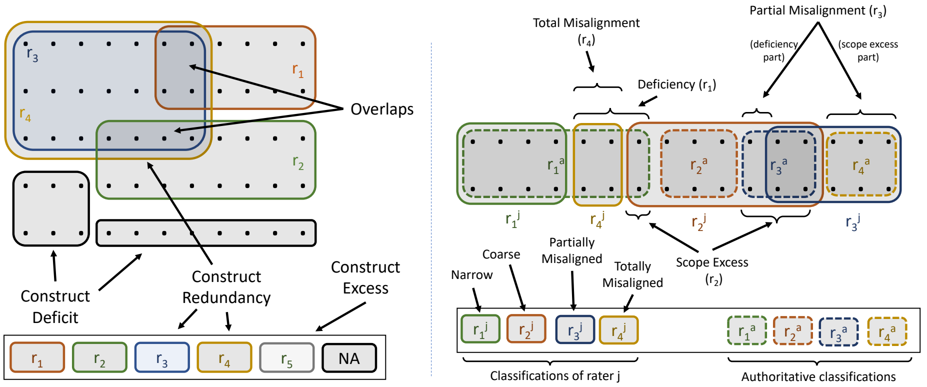

Fig. 1. Rater-based assessments (left) and accuracy (right). Inclusion of dots to solid-lined frames represents frequent inclusion of subjects to the corresponding extensions.

K. Like the other raters, she forms her own authoritative extensions $X_{p_a}(r)$ for each concept term, based on V_D and E. These extensions can be seen as exemplifications of **K**. If most other raters develop extensions that are in agreement with the designer's, it can be empirically argued that the language is conducive to the sharing of **K** between designers and users.

As above, we are interested in indications of imperfect communication of **K**. Considering the sets $X_{p_i}(r)$ of a given rater p_i and the authoritative set $X_{p_a}(r)$:

Perfect Alignment: When $X_{p_i}(r) = X_{p_a}(r)$ the authoritative and the rater's concept are understood to be perfectly aligned.
In our running example of Table 1, let p_1 be the authoritative judge and p_2 a community rater. There is no term r for which $X_{p_1}(r) = X_{p_2}(r)$. Hence there is no occurrence of perfect alignment.

Construct Coarseness: When $X_{p_i}(r) \supset X_{p_a}(r)$ the choice of term r for the concept is too coarse, i.e. evokes an extension that is broader than the concept it is meant to represent. The difference $X_{p_i}(r) \backslash X_{p_a}(r)$ is the *scope excess* of term r with respect to the concept it represents (not to be confused with construct excess).
In our example X_{p_2}(*"wants"*) $\supset X_{p_1}$(*"wants"*), i.e., the term *"wants"* evokes a broader interpretation than what the designer (p_1) expected.

Construct Fineness: When $X_{p_i}(r) \subset X_{p_a}(r)$ the choice of term r for the concept is too narrow, i.e. evokes an extension that excludes elements that the concept it is meant to represent. The difference $X_{p_a}(r) \setminus X_{p_i}(r)$ is the *deficiency* of term r with respect to the concept it is designed to represent.
In our example X_{p_2}(*"desires"*) $\subset X_{p_1}$(*"desires"*), i.e., the term *"desires"* evokes a more restricted set of interpretations than what the designers thought.

Partial and Total Misalignment: When both $X_{p_i}(r) \setminus X_{p_a}(r)$ and $X_{p_a}(r) \setminus X_{p_i}(r)$ are non-empty then the term and the concept are misaligned in a less specific sense. Such misalignment is total when $X_{p_a}(r) \cap X_{p_i}(r) = \emptyset$.
In the example, there is a clear misalignment for each of the terms *"goal"* and *"objective"*, due to, in this case, the overlap between the terms.

Figure 1 (right), offers a schematic showing the logic of the above constructs. Note that the constructs compare the authoritative with the output of one rater. Practical operationalizations must appropriately express the measures in terms of statistics from the output of multiple raters, as we demonstrate below.

4 Application

4.1 Overview and Data Collection

We now present a demonstration of how the empirical constructs developed above can be used to analyze a language. We base the application on real data collected in the context of our earlier experimental study [15] which are here updated and augmented with additional simulated values. An extended presentation can be found in our accompanying report [16] including code snippets, instrument templates, and description examples that can be used for studies with the same or different languages.

The real data were collected in an experiment in which a goal modeling language with concept terms $V_{\mathfrak{R}} = \{$ *"goal"*, *"task"*, *"quality"*, *"belief"*$\}$ was evaluated. Twenty (20) Mechanical Turk workers with a North American bachelor's/college degree, were invited as a proxy for a sample of real language users. They first watched videos that presented the language through informal definitions and examples. Then, four different fictional scenarios were presented in textual form (~250 words each), each corresponding to a description $e \in E$. Beneath each scenario, a set of domain elements V_D mentioned in the scenario were presented – representing a domain of discourse D. Only single elements $d \in V_D$ were presented, hence $\mathbf{D} = V_D$. For each element, the participants were asked to pick one and only one concept term r from $V_{\mathfrak{R}}$ that best describes it. According to what we discussed, the participant response is equivalent to a classification of the subject (d, e) in r, where d is now unary.

To demonstrate the additional empirical constructs we present here, the data was subsequently altered to simulate the following hypothetical conditions. Firstly, a number of elements representing the concept *actor* were part of the experimental prompts, and a term for such actors with the name *"principal"* is added to the language. We assume that the term largely (prob. = 0.9) does not evoke the concept *actor*. Secondly, a *"None of the Above"* (NA) option was included in the options, mentioned henceforth as $r_{\mathtt{NA}}$. We alter the data assuming that if such an option were presented, it would occasionally randomly appear in place of other ratings (prob. = 0.05) and it would be the predominant (prob. = 0.8) response for *actor* instances given the supposed obscurity of *"principal"*. The third hypothetical condition is that in place of the *"belief"* term two terms *"assumption"* and *"assertion"* were part of the vocabulary. To simulate indistinguishability between the two, all *"belief"* ratings are replaced by a random choice of one of those two new terms. We call this initial language $L2$. Given the above manipulations, the data should be indicative of two language problems: (i) a sub-optimal term is used to represent *actor* and (ii) two constructs are overlapping in a way that one of them is redundant.

4.2 Construct Operationalizations

To perform the analysis we first generate concrete operationalizations according to the above data collection method. Let function $n : P \times (\mathbf{D} \times E) \times V_{\Re} \mapsto \{0, 1\}$, be $n(p, s, r) = 1$ if rater p has classified $s = (d, e)$ under r, and $n(p, s, r) = 0$ otherwise; $p \in P, s \in (\mathbf{D} \times E), r \in V_{\Re}$. Denote the marginal sums as, e.g., $n(\cdot, s, r) = \sum_{p \in P} n(p, s, r)$ and likewise for s, r and combinations. Then:
Construct Deficit. We measure construct deficit by calculating the relative proportion of NA responses per element and then identifying elements where such is maximum. Hence, letting $(d, \cdot)$ be subjects of d, the larger the following value the more the evidence for construct deficit of the vocabulary $V_{\Re}$:

$$\mathbf{inc}(V_{\Re}) = \max_{d \in D}\{\frac{n(\cdot, (d, \cdot), r_{\texttt{NA}})}{n(\cdot, (d, \cdot), \cdot)}\}$$

Construct Excess. Let $U(r) = \{n(\cdot, s_1, r), n(\cdot, s_2, r), \ldots\}$ be the set of total classifications each subject s received under r – each, note, is bounded by $|P|$.

Values of the metric below that are closer to 1 indicate construct excess:

$$\mathbf{exc}(r) = 1 - \max[U(r)]/|P|$$

Construct Redundancy. We calculate overlap between r_1 and r_2 on the basis of pairwise disagreements involving the two concepts over the maximum such disagreements can possibly be:

$$ov(s, r_1, r_2) = \frac{n(\cdot, s, r_1) \times n(\cdot, s, r_2)}{\lfloor n(\cdot, s, \cdot)/2 \rfloor \times \lceil n(\cdot, s, \cdot)/2 \rceil}$$

Let $O(r, r') = \{ov(s, r, r') \mid s \in \mathbf{D} \times E\}$ be the set of overlap measures between r and r' over all subjects. Construct redundancy for r can then be measured by:

$$\mathbf{rdn}(r) = \max_{r' \in \mathbf{V}_{\Re} \setminus \{r\}} \{\min[O(r, r')]\}$$

i.e., the maximum overlap exhibited in comparison to every other construct, measured as the minimum of the elementary overlaps that occurred between r and the other construct. To exclude outliers, in all above constructs, percentiles can be used instead of min (in redundancy) and max (in deficit, excess).
Alignment to Authoritative. Given the set of authoritative ratings, we can now define three functions:

- $acc(p, s, r) = \{\mathbf{1}$ if $n(p_a, s, r) = 1$ and $n(p, s, r) = 1$, $\mathbf{0}$ otherwise$\}$
- $def(p, s, r) = \{\mathbf{1}$ if $n(p_a, s, r) = 1$ and $n(p, s, r) = 0$, $\mathbf{0}$ otherwise$\}$
- $exc(p, s, r) = \{\mathbf{1}$ if $n(p_a, s, r) = 0$ and $n(p, s, r) = 1$, $\mathbf{0}$ otherwise$\}$

The marginal totals as per the above notation $acc(\cdot, \cdot, r)$, $def(\cdot, \cdot, r)$, and $exc(\cdot, \cdot, r)$ offer a measure of the *accuracy*, *deficiency* and *scope excess* of a given term vis-à-vis its corresponding concept. The numbers can be used to develop Euler diagrams for visualizing the quality and level of misalignment. An extended discussion on the development of the operationalizations from the empirical constructs introduced earlier is included in our technical report [16].

4.3 Analysis

Let us now explore the output of the metrics given the data we constructed. In the bottom of Figure 2 some indications are shown for language $L2$. The

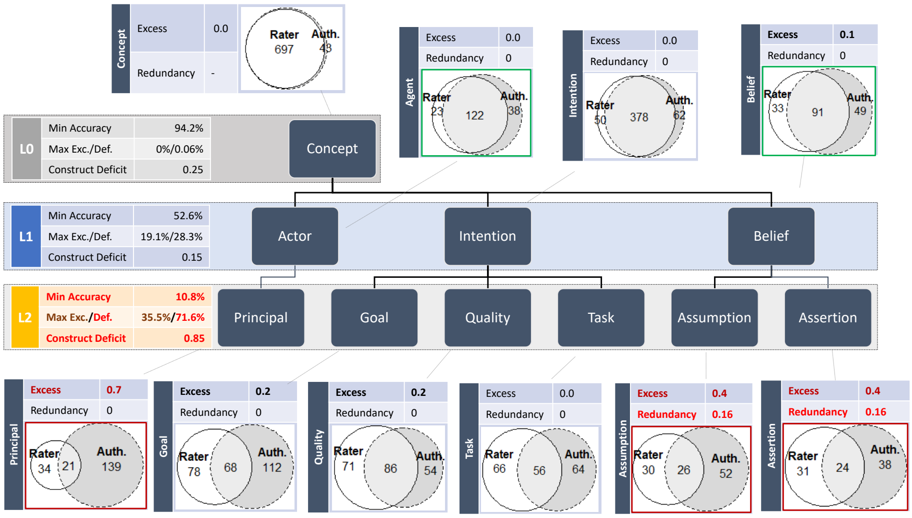

Fig. 2. Analysis of two languages.

construct excess indices are 0.7, 0.4, and 0.4, for *"principal"*, *"assumption"* and *"assertion"*, indicating possible excess issues with each construct. Furthermore, redundancy is zero everywhere except for *"assumption"* and *"assertion"*, meaning a possible overlap between the two. A look at the accuracy Euler diagrams shows that accuracy, i.e., the intersection of the rater and authoritative circles, is very low in those constructs. For *"principal"*, the deficiency of the construct, i.e. all the ratings it should attract but it did not, is also notable. On the table on the left, general measures about the language can be seen, including the minimum accuracy observed across all constructs, and the maximum scope excess and deficiency measures observed for the language. The construct deficit index of $L2$ is 0.85, meaning that some elements are not represented in the conceptualization evoked by the language.

After observing the results, assume that we decide to engage in corrective measures, resulting in language $L1$, as follows. What *"principal"* used to represent has now been renamed as *"actor"*. To simulate raters now successfully recognizing the *actor* concept, the corresponding elements are classified to that construct rather than $r_{\mathtt{NA}}$ (simulated with prob. = 0.8). Further, the constructs *"assertion"* and *"assumption"* are merged into *"belief"*; the corresponding ratings are reverted to the original. Finally, constructs *"goal"*, *"quality"* and *"task"* are replaced by a new construct *"intention"*. Assuming that raters who classified a subject under one of the three original terms, would have classified the same subject under *"intention"*, the corresponding classifications are replaced accordingly in the data. We can see in Figure 2 that for $L1$, the excess and redundancy measures are now normal, and the accuracies have improved.

Figure 2 also shows how lowering the granularity [9] of the concepts may result in an improvement of the proposed metrics. At the extreme, language

$L0$ includes only one concept, called *"concept"* allowing for limited room for disagreement and inaccuracies. However, such language as $L0$ may lack the expressiveness needed, and the construct may suffer from *construct overload* [25]. As opposed to the other quality characteristics, measuring overload by ratings from a given language alone is difficult. Rather, when refinement of the language is attempted and the result is a language that performs well in all other aspects, then we can hypothesize the presence of remediable construct overload in the original language.

4.4 Validity Threats

The above analysis is a demonstration of the metrics based on data that has been simulated to exactly exhibit their merits. In studies with real data, experimenters need to be mindful of some validity threats and limitations.

In terms of *external validity*, that is, the generalizability of an analysis, the metrics are as good indicators as the representativeness of the world descriptions and domain elements. Sampling that consistently leaves out a class of domain phenomena, will result in false construct redundancy or construct excess indications. Inflated construct deficit indications may also emerge when phenomena that are irrelevant to the language are included in the samples. Note, further, that the choice of a Mechanical Turk sample in our study was possible for demonstration purposes due to the familiarity of the broad population with the concepts considered. However, for evaluating a language against a specific group of prospective users, external validity requires the selection of a representative sample from that exact group. Further, while individual differences in terms of linguistic ability and expertise will affect the outcome of a language evaluation, if the rater sample is representative, then whatever variability and issues emerge will still reflect the quality of the language for the given user group. The main concern regarding *internal validity* is the relationship between the metrics and the pathology they indicate. Although the constructs are direct consequences of the pathology definitions, the question of whether they constitute necessary and/or sufficient evidence for the presence of the pathology is a matter for further investigation. One of the enablers of such correspondence is proper operationalization of the metric including both the statistical instantiations of the constructs, to control for, e.g., chance responses, and the data collection instruments.

Overall, empirical evaluation of our framework is largely interpreted into studying the quality of the instruments developed for a specific language in question, including questionnaire format, training material, descriptions, and domain elements. Established techniques for instrument quality assessment can help with this task. These include measuring retest reliability, which establishes if the same rater produces the same rating at different times, and inter-rater reliability, which refers to the agreement between raters. For the latter, however, a benchmark language and instrument with known good quality, as per, e.g., expert opinion, would need to be used. In this way, possible disagreements can be attributed to the instrument or process rather than to the language. Likewise, specific issues (construct redundancies, excesses, etc.) can be introduced to the language by experts, for checking if the instruments accurately detect them – a method we simulated in our study. Finally, languages with large vocabularies imply longer and more complex rater tasks. When this appears to threaten experimental task integrity, evaluation can take place in a piecemeal fashion whereby either different groups of raters are given different domain elements and descriptions or different, possibly semantically related, subsets of terms are evaluated separately – at the expense of not capturing issues that span across subsets. These strategies and when they are necessary are yet to be investigated.

5 Related Work

Several efforts for empirically evaluating modeling languages have been proposed in the literature. One line of work concerns the identification of language quality dimensions that are subject to evaluation [13,21]. The notions of *comprehensibility appropriateness* and *domain appropriateness* are the most closely relevant to the Wand and Weber framework adopted here. Similar notions include *semantic transparency* and *semiotic clarity*, as discussed by Moody [20].

Empirical efforts for assessing model understandability have also been reported. Houy et al. [11] survey empirical studies that evaluate the particular construct for various kinds of models including entity, class, and process models. Requirements goal models have also been a focus of such investigation [2,5,8,10,24]. In our past work, we studied the *intuitive* (i.e. without training) evocation of the meaning of a language construct via observing inferences participants perform with the construct [1,14,17,18]. However, most of these empirical efforts focus on diagrammatic constructs (boxes, lines, icons) and their visual efficiency, rather than on the choice of terms.

Work focused on terms and concepts can be found in the area of ontology engineering. Annotation of text as a terminology building and evaluation step has been proposed [3,26]. Measures of inter-rater agreement [12] to attain *semantic agreement* can be applied in such exercises [22]. Ontology learning techniques also have components that are relevant to our proposal [3]. An important process in ontology learning [3,26] is *term extraction*, i.e., the identification of terms that are relevant in the domain – e.g. [19]. Term extraction serves the purpose of supporting domain appropriateness in that terms are grounded on “true” discourse taking place in the domain as documented in the texts being processed. Comparatively, our process is geared towards evaluating an existing terminology and characterizing its appropriateness in a way that informs improvement.

Finally, analytical methods can promote the sharedness of an ontological commitment. Developing ontologies [6] allows explication of the commitment through the formulation of properties of the terms within language, e.g., meaning postulates, that are consequences of the commitment. Upper-level ontologies can, further, be used to identify issues with a language meta-model [7]. Empirical analyses are meant to complement such approaches and to also measure the extent to which a language is learned by the community of practitioners.

6 Conclusions and Future Work

We presented a framework for empirically measuring the appropriateness of vocabulary choices for conceptual modeling languages. The framework is based on the measurement of the degree of sharedness of the ontological commitment of the language via observing how experimental participants map descriptions of possible worlds to extensions of the vocabulary terms. A set of empirical constructs are defined for characterizing the resulting mappings in terms of specific pathologies of the vocabulary choice as per an established model. The constructs allow different concrete operationalizations that fit the needs of specific data collection techniques. We demonstrated the utility of the framework over a hypothetical language design scenario using a mix of real and simulated data.

There are several opportunities for further consolidating and extending our framework, in addition to empirical evaluation suggestions mentioned above. These include analytically and empirically studying possible operationalizations of our proposed constructs with respect to exhibiting statistical properties suitable for generalizations and comparisons. Further, experiments with various languages need to be conducted both for validation and for the establishment of community norms/baselines as is commonly the case with standardized empirical instruments – e.g., what levels of "excess" or "deficit" are common.

References

1. Alothman, N., Zhian, M., Liaskos, S.: User Perception of Numeric Contribution Semantics for Goal Models: an Exploratory Experiment. In: Proc. of the 36th Int. Conf. on Conceptual Modeling (ER 2017). pp. 451–465. Xi'an, China (2017)
2. Caire, P., Genon, N., Heymans, P., Moody, D.L.: Visual notation design 2.0: Towards user comprehensible requirements engineering notations. In: Proc. of the 21st IEEE Int. Req. Eng. Conf. (RE'13). pp. 115–124. Rio de Janeiro, Brasil (2013)
3. 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, Berlin, Heidelberg (2009)
4. Dalpiaz, F., Franch, X., Horkoff, J.: iStar 2.0 Language Guide. The Computing Research Repository (CoRR) **abs/1605.0** (2016), `http://arxiv.org/abs/1605.07767`
5. Estrada, H., Rebollar, A.M., Pastor, O., Mylopoulos, J.: An Empirical Evaluation of the i* Framework in a Model-Based Software Generation Environment. In: Proc. of the 18th International Conference on Advanced Information Systems Engineering (CAiSE'06). pp. 513–527. Luxembourg, Luxembourg (2006)
6. Guarino, N., Oberle, D., Staab, S.: What Is an Ontology? In: Handbook on Ontologies, pp. 1–17. Springer (2009)
7. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. Ph.D. thesis, University of Twente (2005)
8. 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)

9. Henderson-Sellers, B., Gonzalez-Perez, C.: Granularity in Conceptual Modelling: Application to Metamodels. In: Proc. of the 29th International Conference on Conceptual Modeling (ER 2010). pp. 219–232. Vancouver, Canada (2010)
10. Horkoff, J., Yu, E.: Finding solutions in goal models: an interactive backward reasoning approach. In: Proc. of the 29th International Conference on Conceptual modeling (ER 2010). pp. 59–75. Vancouver, Canada (2010)
11. 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). pp. 64–77. Florence, Italy (2012)
12. Krippendorff, K.: Content Analysis: An Introduction to it Methodology. SAGE (2004)
13. Krogstie, J.: Model-Based Development and Evolution of Information Systems. Springer (2012)
14. Liaskos, S., Dundjerovic, T., Gabriel, G.: Comparing Alternative Goal Model Visualizations for Decision Making: an Exploratory Experiment. In: Proc. of the 33rd ACM Symp. on Applied Computing (SAC'18). pp. 1272–1281. Pau, France (2018)
15. Liaskos, S., Jaouhar, I.: Towards a framework for empirical measurement of conceptualization qualities. In: Proc. of the 39th International Conference on Conceptual Modeling (ER 2020). pp. 512–522. Vienna, Austria (2020)
16. Liaskos, S., Mylopoulos, J., Khan, S.M.: Replication Data for: Empirically Evaluating the Semantic Qualities of Language Vocabularies. Scholars Portal Dataverse (2021). https://doi.org/10.5683/SP2/H4BHLT
17. Liaskos, S., Ronse, A., Zhian, M.: Assessing the Intuitiveness of Qualitative Contribution Relationships in Goal Models: an Exploratory Experiment. In: Proc. of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'17). pp. 466–471. Toronto, Canada (2017)
18. Liaskos, S., Tambosi, W.: Factors Affecting Comprehension of Contribution Links in Goal Models: An Experiment. In: Proc. of the 38th International Conference on Conceptual Modeling (ER 2019). pp. 525–539. Salvador, Brazil (2019)
19. Medelyan, O., Witten, I.H.: Thesaurus-based index term extraction for agricultural documents. In: Proc. of the 2005 EFITA/WCCA Joint Congress on IT in Agriculture. pp. 1122–1129. EFITA/WICCA, Vila Real, Portugal (2005)
20. Moody, D.L.: The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. IEEE Transactions on Software Engineering (TSE) **35**(6), 756–779 (2009)
21. Nelson, H.J., Poels, G., Genero, M., Piattini, M.: A conceptual modeling quality framework. Software Quality Journal (20), 201–228 (2012)
22. 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 US, Boston, MA (2007)
23. for Requirements Engineering, S.M. (ed.): Yu, Eric and Giorgini, Paolo and Maiden, Neil and Mylopoulos, John. MIT Press (2011)
24. Santos, M., Gralha, C., Goulão, M., Araújo, J.: Increasing the Semantic Transparency of the KAOS Goal Model Concrete Syntax. In: Proc. of the 37th International Conf. on Conceptual Modeling (ER 2018). pp. 424–439. Xi'an, China (2018)
25. Wand, Y., Weber, R.: On the ontological expressiveness of information systems analysis and design grammars. Inf. Systems Journal **3**(4), 217–237 (1993)
26. Wong, W., Liu, W., Bennamoun, M.: Ontology Learning from Text: A Look Back and into the Future. ACM Computing Surveys **44**(4) (sep 2012)