Evaluating the Completeness and Granularity of Functional Requirements Specifications: A Controlled Experiment

Sergio España¹, Nelly Condori-Fernández¹, Arturo González², Óscar Pastor¹

¹Centro de Investigación en Métodos de Producción de Software
²Departamento de Sistemas Informáticos y Computación, ¹,²Universidad Politécnica de Valencia
{sergio.espana, nelly, opastor}@pros.upv.es, agdelrio@dsic.upv.es

Abstract

Requirements Engineering (RE) is a relatively young discipline, and still many advances have been achieved during the last decades. In particular, numerous RE methods have been proposed. However, there is a growing concern for empirical validations that assess RE proposals and statements. This paper is related to the evaluation of the quality of functional requirements specifications, focusing on completeness and granularity. To do this, several concepts related to conceptual model quality are presented; these concepts lead to the definition of metrics that allow measuring certain aspects of a requirements model quality (e.g. degree of functional encapsulations completeness with respect to a reference model, number of functional fragmentation errors). A laboratory experiment with master students has been carried out, in order to compare (using the proposed metrics) two RE approaches; namely, Use Cases and Communication Analysis. Results indicate greater quality (in terms of completeness and granularity) when Communication Analysis guidelines are followed. Moreover, interesting issues arise from experimental results, which invite further research.

1. Introduction

Requirements Engineering (RE), in spite of being a young discipline, has achieved an ever growing body of knowledge. Numerous RE methods have been proposed over the last decades. Most authors act as designers and propose new RE methods, while few act as real researchers validating that their (or other’s) proposals actually improve RE practice [1]. Empirical evaluations are a strong need in the area [2]. A grand challenge for the field is to develop an understanding of which software methods work better and why [3].

In order to carry out method evaluations and comparisons, many evaluation techniques are available [4]. Some techniques are theoretical (e.g. ontological analysis) and others are empirical (e.g. laboratory experiments); each type having its strengths and weaknesses. Laboratory experiments provide high level of control and they are powerful in determining causality (strong internal validity), while their artificial environment compromises the generalisation of the results (external validity is their weakness) [4][5].

This paper presents a laboratory experiment that evaluates and compares two RE methods; namely Use Cases [6] and Communication Analysis [7]. Strictly speaking, the products of the following Requirements Specification (RS) techniques¹ are compared: Use Case Diagrams and Communicative Events Diagrams (and their corresponding textual descriptions). These RS techniques allow the graphical and textual specification of what the literature usually refers to as functional requirements. The focus of interest is put on two aspects of conceptual model² quality: functional completeness and appropriate granularity.

Many authors have theorised and empirically validated conceptual model completeness (the degree to which a model specifies all the relevant statements of a domain) [8][9]. However, the approach often consists of a reviewer rating completeness on a Likert scale (or similar), and the procedure for assigning the rating depends on a subjective judgement [8][10][11]. We propose an approach based on the comparison of the reviewed model with a reference model that is built and agreed by an expert modelling committee. The metrics and the measuring procedure are also defined.

¹ We refer as method to a systematic way of working by which one can obtain a desired result. We refer as technique to a recipe for obtaining a certain result. Methods contain techniques [27].
² Requirements specifications are considered conceptual models that represent a problem domain and user needs [27].
Granularity is a much less investigated quality aspect, although it is an issue that has provoked debate, particularly with regard to functional requirements such as use cases [12][13]. Kulak and Guiney define use case granularity as the relative scope of individual use cases compared to the application’s scope [13]. In a more general sense, granularity measures the size of an encapsulation (a systemic notion). When modelling, the analyst relies on methodological guidelines in order to encapsulate concepts in modelling primitives. We refer to the criteria that allow determining granularity as unity criteria [14]. This paper proposes an evaluation of granularity that goes beyond previous proposals since sound unity criteria for business process modelling are taken as reference and precise metrics for granularity errors are defined.

The contributions of the paper are the following:

- The paper discusses the concepts of completeness and granularity from a semantic point of view and grounds them in sound theory.
- The concepts of completeness and granularity are operationalised by proposing metrics for their quantification. Some metrics extend previous approaches; other metrics adopt a novel approach.
- The paper presents an experiment that compares two functional RS techniques (Use Case Diagrams and Communicative Events Diagrams), using the above-mentioned concepts and metrics. Although we focus on two particular methods, the proposed strategy is general enough to fit the evaluation of other methods.
- The paper is structured as follows. Section 2 reviews previous quality frameworks, proposes concepts and metrics related to completeness and granularity of functional RS, and reviews previous evaluations of RS techniques. Section 3 overviews the RS techniques evaluated in the laboratory experiment; namely, Use Case Diagram and Communicative Events Diagram. Section 4 presents the experimental planning. Section 5 analyses the results of the experiment. Section 6 discusses the validity of the results. Section 7 presents conclusions and future work.

2. Comparison of Functional Requirements Specification Techniques

2.1. Conceptual model quality frameworks

Lindland et al. [9] present a conceptual model quality framework that it is founded on Semiotics and Linguistics. Three types of model quality are assumed and absolute goals are defined (see Figure 1).

- **Syntactic quality**. The degree to which the model adheres to the language rules. Syntactic errors and deviations from the rules decrease syntactic quality.
- **Semantic quality**. The degree to which the model represents the domain. The more similar the model and the domain, the better the semantic quality.
- **Pragmatic quality**. The degree to which the model is correctly interpreted by its audience. The less misunderstanding, the better the pragmatic quality.

![Conceptual model quality framework][1]

Figure 1. Conceptual model quality framework [9]

Other frameworks have been proposed in the literature. Yadav et al. [8] classify criteria for the comparison of RE methods in four categories: syntactic, semantic, usability, and communicating ability. Davis et al. [15] explore the concept of RS quality and consider completeness, but acknowledge that the proposed metrics are “difficult to measure”. Pohl [16] develops a framework for RE with three dimensions - with their respective goals (specification-formally represented, representation-complete and agreement-agreed). In [17], the framework by Lindland et al. is extended by considering additional levels of the semiotic ladder [18]. Moody et al. [19] present a quality framework for data models oriented towards practice. Schuette et al. [20] present a framework of principles that improve the quality of information models by reducing subjectivism in the modelling process. State of the art reviews and framework comparisons can be found in [21], [22], [23] and [24].

We have chosen the framework by Lindland et al. as a point of departure for our research for it has been acknowledged as a reference framework [17][24] and it has been empirically tested [10][11].

2.2. Completeness and granularity from a semantic perspective: concepts and metrics

Lindland et al. [9] define that a model (M) has achieved semantic completeness if it contains all the statements about the domain (D) that are correct and relevant. That is, \(DM=\emptyset\). However, except for extremely simple and highly inter-subjectively agreed domains, total completeness cannot be achieved (due to resource restrictions). Hence, they relax the
completeness goal by applying the notion of feasibility. Feasibility introduces a trade-off between the benefits and drawbacks for achieving a given model quality.

A model has achieved feasible completeness when there is no relevant statement about the domain, not yet included in the model, such that the additional benefit to the conceptual model from including the relevant statement exceeds the drawbacks of including it. That is, \( D \setminus M = S \neq \emptyset \), where \( S \) is the set of correct and relevant statements not yet in the model.

However, from a constructivist stance (such as in [17]), to determine whether there is a relevant statement of the domain not yet included in the model, one must first conceptualise the domain. This conceptual model of reference (\( M_r \)) needs not be written but at least it must exist in the reviewer’s mind.

Also, Lindland et al. recommend adapting and refining the framework depending on the method being evaluated, as well as operationalising the framework by proposing metrics. For instance, they propose to decompose feasible completeness into feasible functional completeness, feasible non-functional completeness, etc. A germinal definition states that a model has achieved feasible functional completeness whenever all relevant functional requirements that are worth being specified have been included in the model (adapted from [9]). We further develop this idea by refining it, by explicitly considering the existence of a reference model, and by proposing metrics.

When Lindland et al. propose refining the framework they consider the existence of different types of statements about the domain, what leads to building models with multiple views. Each view can itself be considered a model. For instance, if the domain is considered to contain two types of statements (those about functionality \(-FD-\) and those about qualities and restrictions \(-NFD-\)), then the requirements model \( M \) will have two views: a functional requirements model \( FM \) and a non-functional requirements model \( NFM \); that is, \( D=FD\cup NFD \) and \( M=FM\cup NFM \). Feasible functional completeness can now be formally defined as \( FDFM=FS\neq\emptyset \), where \( FS \) is the set of functional requirements not yet in the model (but which are not worth the effort to be included).

Feasible functional completeness can be further refined if a specific type of domain is considered (e.g., information systems). Research about the essence of information systems has lead to the definition of sound conceptual frameworks. An information system (IS) is a socio-technical system that supports organisational communications [25]. FRISCO report [18] laid the foundations of the area, upon which other researchers have built theories [26]. Based on this conception of ISs, two major abstract modelling primitives for functional RS are identified; namely functional encapsulations and linked communications.

- Wieringa defines function as a service provided by the IS to its environment; it is also considered to be an encapsulation of a useful behaviour of the system [27]. We therefore refer as functional encapsulations to IS functions, in order to highlight the importance of determining the boundaries of the encapsulation.

- We refer as linked communication to the message conveyance that is triggered by the occurrence of an event (the use or activation of a function) and by which the IS informs an actor of this occurrence.

![Figure 2. Abstraction of the primitives of ISs functional requirements models](image)

This way, it is considered that a functional requirements model is composed of (at least) a set of functions (\( F \)) and a set of linked communications (\( LC \)). Then, \( FM=FS\cup LC \). Figure 2 shows an example of these abstract primitives and Section 3 establishes their correspondence with the evaluated RS techniques.

Furthermore, a constructivist approach takes us to define completeness with respect to a reference model. In industrial settings, a reference model may be impractical; customer reviews are essential [15]. In experimental settings, we propose that an expert modelling committee analyse a given domain and agree a model that strictly follows best practices in modelling\(^3\). Some best practices depend on the modelling technique (e.g., correct use of use case includes and extends relations [6]) whereas others are independent (e.g., RS should provide an external view of the system [28]). The reference model is defined as \( M_r=FM_r\cup NFM_r \), and the reference functional requirements model is defined as \( FM_r=Fr\cup LC_r \). It is assumed that \( FM_r \) contains all relevant functional requirements and linked communications.

\(^3\) This agreement may involve debate and several iterations.
Measurable quality goals are defined in terms of the reference model (see Figure 3):

- **Functional encapsulations completeness with respect to (w.r.t.) a reference model.** All functional requirements specified in the reference model have been specified in the model. That is, $F \cap F' = \emptyset$. Note that the reference model substitutes the domain in the comparison.

This goal is related to the degree with which the reviewed model contains the functional encapsulations (i.e. use cases, communicative events) included in the reference model. A metric for this goal is the degree of functional encapsulations completeness w.r.t. a reference model, which is defined as $\deg FEC = |F \cap F'|$.

- **Linked communications completeness w.r.t. a reference model.** All linked communications specified in the reference model have been specified in the model. That is, $LC \cap LC' = \emptyset$.

This goal is related to the degree with which the reviewed model contains the linked communications (i.e. communications triggered by the occurrence of a use case or a communicative event) specified in the reference model. A metric for this goal is the degree of linked communications completeness w.r.t. a reference model, which is defined as $\deg LCC = |LC \cap LC'|$.

Assuming that it is feasible to build the reference model, both goals are feasible quality goals.

![Figure 3. Model completeness evaluation with respect to a reference model](image)

With regard to granularity, we claim that it is a quality of requirements models that has not been sufficiently investigated. We refer as **unity criteria** to the norms that guide the identification of complex concepts (e.g. use cases) and the encapsulation of their components (i.e. the flow of actions a use case is composed of); therefore, unity criteria determine the granularity of encapsulations [14]. Data models unity criteria are frequently based on the notion of identification; there is much consensus in this area. However, unity criteria for functional models (e.g. use case models) are far from being widely agreed. We argue that this lack of agreement leads to the proliferation of methodological guidelines that avoid the thorny issue of granularity or, in many cases, they define simplistic unity criteria (e.g. the **twenty use cases per system rule of thumb** [29], the **one person, one sitting test** [6]). Consequences are inconsistencies in modelling practice, heterogeneous model granularity and the need for gurus to whom consult [30].

In [14] we unfold the notion of unity criteria and we propose unity criteria for business process modelling. Industrial practice has shown us that the criteria reduce subjectivity in the encapsulation of business processes.

A model is considered to have an **appropriate granularity** with respect to a given unity criteria when the encapsulations of the model (typically under the form of modelling primitives) are conform to the unity criteria. Taking a given set of unity criteria it is possible to identify granularity errors in a model (see Figure 4). If the reference model conforms to the unity criteria, it aids the review; but it is not strictly required.

![Figure 4. Model granularity evaluation with respect to unity criteria](image)

For functional requirements models we propose identifying the following granularity errors:

- **Functional fragmentation error.** This error is the result of modelling two or more functional encapsulations for a part of the domain which, according to the given unity criteria, should have been modelled as only one encapsulation. For instance, two or more use cases of the reviewed model correspond to one use case of the reference model. It is measured as a variable named $\text{errFra}$.  

- **Functional aggregation error.** This error is the result of modelling certain part of the domain as one functional encapsulation when, according to the unity criteria, the phenomena should be modelled using two or more encapsulations. For instance, two or more use cases of the reference model are modelled as only one use case in the reviewed model. It is measured as a variable named $\text{errAgg}$.  

Therefore, a model can be said to achieve appropriate granularity whenever there are no granularity errors. That is, $errFrac=0$ and $errAgr=0$.

2.3. Previous empirical evaluations and comparisons of RS techniques

Some empirical works assess the quality of functional requirements models. Special interest has been placed in evaluating use case specification guidelines. In regard to completeness, two approaches can be distinguished: assessing the completeness of a (single) use case description, rating the completeness of a use case model based in the reviewer’s judgement, and measuring the size of the use case model.

Some works focus on a single use case and analyse its detailed textual description. For instance, an experiment by Ben Achour et al. [31] lays emphasis on the specification of the use case flow of actions. Cox & Phalp [32] replicate the previous experiment and extend the marking scheme with subjective metrics. Instead, as explained in Section 2.2, we are interested in assessing the whole functional RS.

Other works focus on the whole use case model but rate completeness in terms of a value judgement. Yadav et al. [8] propose assessing completeness by having an expert review committee rate each model on a 1 to 7 scale, based on judgement. In experiments by Moody et al. [10][11] quality ratings are also given on a 7 point Likert scale, from 1 (poor) to 7 (excellent). We advocate using metrics rather than ratings.

In [33], the authors use a 0 to 3 scale to rate several properties of the model (mainly related to correctness). In regard to completeness, the paper states that “the number of identified actors and use cases [...] indicate quality of the guidelines – the higher number, the better quality”. We believe that this statement is only sensible if just valid actors and use cases are counted and if use cases have an appropriate granularity, but the paper does not give a more detailed explanation.

Fortuna [34] describes an experiment that considers functional coverage and granularity homogeneity in order to validate unity criteria for use cases; however, the procedure is not sufficiently described to learn lessons from it.

3. Use Cases and Communication Analysis

We chose to compare Use Cases and Communication Analysis for the following reasons:
- Both proposals are applicable to the common domain of information systems (ISs) development.
- The authors have practical experience in using both.
- One of the proposals (Use Cases) is widely known and it is used worldwide. The other proposal (Communication Analysis) is less known. However, several Spanish companies apply it successfully; this was a major reason to select it for the study.

![Figure 5. Use Case Model fragment](image1)

**Figure 5.** Use Case Model fragment

Use Cases were proposed by Jacobson [35] and later revised by many authors. “A use case is a collection of possible sequences of interactions between the system under discussion and its external actors, related to a particular goal” [36]. Use Cases have been used in many industrial projects [37] but they have also generated strong debates on their usefulness [38]. With regard to the relation between the Use Case Diagram technique and the abstract primitives see Figure 5. Functional encapsulations correspond to use cases. Linked communications typically appear in the use case description (e.g. as a step in the flow of actions), but may as well appear as a relation between a use case and an actor. Many works propose guidelines for use case modelling [39][40]; guidelines by Cockburn [6] were chosen for the experiment because he explicitly proposes unity criteria that are based on user goals.

![Figure 6. Communicative Event Diagram fragment](image2)

**Figure 6.** Communicative Event Diagram fragment

Communication Analysis is a software development method that proposes undertaking ISs development from a communicational perspective [7]. The method stems from ISs foundations research [26] and it evolves by means of the collaboration between industry and academia. We choose to assess the Communicative Event Diagram because it is comparable to the Use Case Diagram. This technique is intended to describe business processes from a...
communicational perspective. With regard to the relation between the notation and the abstract primitives see Figure 6. Functional encapsulations are communicative events and linked communications are outgoing communicative interactions [7].

4. Experimental Planning

The goal of our experiment, according to the Goal/Question/Metric template [41], is to analyze functional requirements specifications (RS) with the purpose of carrying out a comparative evaluation of RS techniques with respect to their completeness and granularity from the viewpoint of the researcher in the context of Computer Science master students. The experiment addresses the following research questions:

- **RQ1:** Will the subjects applying Communication Analysis produce functional RS with higher degree of completeness than the subjects applying Use Cases?
- **RQ2:** Will the subjects applying Communication Analysis produce functional RS with a more appropriate level of granularity than the subjects applying Use Cases?

4.1. Experimental Context

The selected subjects were 36 Computer Science master students enrolled in the 2007-2008 “Conceptual Modelling of Information Systems” course at Universidad Politécnica de Valencia, Spain; visit http://www.fiv.upv.es/default_i.htm for more details on the degree. Participation was anonymous (aliases were used instead of names) and students were ensured that performance would not influence academic marks. The subjects received lectures on Communication Analysis and Use Cases techniques. Two groups were formed according to student’s availability; each subject applied each of both techniques to specify the requirements of a Photography Agency IS. The experiment was carried out in two rounds (see Table 1).

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication Analysis</td>
<td>Group 1 (N=27)</td>
<td>Group 2 (N=7)</td>
</tr>
<tr>
<td>Use Cases</td>
<td>Group 2 (N=8)</td>
<td>Group 1 (N=28)</td>
</tr>
</tbody>
</table>

4.2. Variables

We identified three types of variables [5]:
- Response variables. Functional completeness and granularity were identified as outcomes of the experiment. Section 2.2 defines the metrics: $\text{degFEC}$, $\text{degLCC}$, $\text{errFra}$, and $\text{errAgg}$.
- Factors. The requirements specification technique was identified as a variable that could affect the response variables. Two treatments were considered: (1) Use Cases (mainly Use Case Diagram and textual use case descriptions) and (2) Communication Analysis (mainly Communicative Event Diagram).
- Parameters. Variables that we do not want to influence the experimental results have been fixed: application domain, complexity of the IS (problem statement) and experience using RS techniques.

4.3. Hypotheses

The hypotheses formulated from the research questions defined above are the following:

**Hypothesis 1.**

Null hypothesis, $H_{10}$. Use Cases and Communication Analysis allow obtaining RS with same degree of functional encapsulations completeness.

Alternative hypothesis, $H_{11}$. Communication Analysis allows obtaining RS with greater degree of functional encapsulations completeness than Use Cases.

**Hypothesis 2.**

Null hypothesis, $H_{20}$. Use Cases and Communication Analysis allow obtaining RS with same degree of linked communications completeness.

Alternative hypothesis, $H_{21}$. Communication Analysis allows obtaining RS with greater degree of linked communications completeness than Use Cases.

**Hypothesis 3.**

Null hypothesis, $H_{30}$. Use Cases and Communication Analysis allow obtaining RS with same number of functional fragmentation errors.

Alternative hypothesis, $H_{31}$. Communication Analysis allows obtaining RS with less functional fragmentation errors than Use Cases.

**Hypothesis 4.**

Null hypothesis, $H_{40}$. Use Cases and Communication Analysis allow obtaining RS with same number of functional aggregation errors.

Alternative hypothesis, $H_{41}$. Communication Analysis allows obtaining RS with less functional aggregation errors than Use Cases.

4.4. Experimental Procedure

The experiment was initiated with a demographic questionnaire which showed that the subjects had similar background; for instance, 83% of them had little knowledge about the syntax of Use Cases (score
inferior to 3 points using a 1-5 Likert scale), and none knew about Communication Analysis. Then the subjects were trained on a RS method. During the first round, Group 1 was trained in Communication Analysis, and Group 2 was trained on Use Cases. Then the subjects received a natural language problem statement describing the structure and processes of a Photography Agency. They applied the RS method to specify the needed IS. The second round was analogous but now Group 1 was trained on Use Cases, and Group 2 was trained on Communication Analysis. Each round lasted 8 hours distributed over 4 days.

**Figure 7. Experimental procedure**

### 5. Results Analysis and Interpretation

An expert reviewer used the reference model and the unity criteria to aid him in the correction of the RS models. The measures obtained by the expert reviewer were analyzed. See Table 2 (a suffix is added to the variable name to indicate the technique: CA stands for Communication Analysis, UC stands for Use Cases).

#### Table 2. Descriptive statistics

<table>
<thead>
<tr>
<th>Measure</th>
<th>N</th>
<th>Mean</th>
<th>Std. deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>degFEC_CA</td>
<td>34</td>
<td>0.8100</td>
<td>0.11710</td>
</tr>
<tr>
<td>degFEC_UC</td>
<td>36</td>
<td>0.7564</td>
<td>0.11557</td>
</tr>
<tr>
<td>degLCC_CA</td>
<td>34</td>
<td>0.7500</td>
<td>0.22937</td>
</tr>
<tr>
<td>degLCC_UC</td>
<td>36</td>
<td>0.5046</td>
<td>0.15164</td>
</tr>
<tr>
<td>errFra_CA</td>
<td>34</td>
<td>1.8824</td>
<td>0.97746</td>
</tr>
<tr>
<td>errFra_UC</td>
<td>36</td>
<td>2.1944</td>
<td>1.75368</td>
</tr>
<tr>
<td>errAgg_CA</td>
<td>34</td>
<td>0.5588</td>
<td>0.78591</td>
</tr>
<tr>
<td>errAgg_UC</td>
<td>36</td>
<td>1.2222</td>
<td>0.68080</td>
</tr>
</tbody>
</table>

The degree of functional encapsulations completeness w.r.t. a reference model (degFEC) is 81% for Communication Analysis and 75.64% for Use Cases. A bigger difference appears in the degree of linked communications completeness (degLCC), with up to 75% for Communication Analysis and only 50.46% for Use Cases. With regard to granularity errors, both error measures indicate that subjects applying Communication Analysis perform better than subjects applying Use Cases.

#### 5.1. Analyzing functional Completeness w.r.t. a reference model

By applying the Kolmogorov-Smirnov test, we noted that both degFEC and degLCC measures had a normal distribution (p>0.5), Paired Sample Test was applied to verify the null hypotheses $H_{10}$ and $H_{20}$.

<table>
<thead>
<tr>
<th>Measure</th>
<th>t</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>degFEC_CA</td>
<td>1.96</td>
<td>0.058</td>
</tr>
<tr>
<td>degLCC_CA</td>
<td>1.96</td>
<td>0.058</td>
</tr>
</tbody>
</table>

As we can see in Table 3, there is a medium significance difference (p=0.05) between the Use Cases and the Communication Analysis techniques, with respect to the degree of functional encapsulations completeness (degFEC). Besides, there is a very high significance difference (p=0.000) with respect to the degree of linked communications completeness w.r.t. a reference model (degLCC). Therefore, $H_{10}$ and $H_{20}$ are refuted with a 95% confidence and the alternative hypotheses $H_{11}$ $H_{21}$ are verified. This means that Communication Analysis allows obtaining RS with greater degree of functional encapsulations and linked communications completeness than Use Cases.

This outcome can be explained by the fact that Communication Analysis methodological guidelines for the identification and specification of system functions follow a more systematic procedure than Use Cases guidelines. Also, Communication Analysis approaches functional requirements specification from a business process perspective; this way, temporal precedence relations of the specified processes facilitate the discovery of missing communicative events, thus contributing to higher completeness. Moreover, the Commutative Event Diagram technique devotes a modelling primitive to linked communications (outgoing communicative interactions) so they are explicitly specified.

#### 5.2. Analyzing appropriate granularity

The Kolmogorov-Smirnov test was applied to normality test of both errFra and errAgg measures. As we note that only errFra measure had a normal distribution (p>0.5), Paired Sample Test was also applied to verify the null hypothesis $H_{30}$. We note in Table 4 that, with respect to fragmentation errors
(errFra), there is not a significant difference (p=0.352) between the Use Cases and the Communication Analysis techniques. Therefore, hypothesis H31 was not corroborated.

By using the Wilcoxon signed-rank non-parametric test for verifying null hypothesis H40, we observe in Table 5 that 20 of 34 RS models using Use Cases show a greater number of functional aggregation errors than using Communication Analysis. This statistical differences presented a high significance level (see Table 6, p=0.002). Therefore, hypothesis H41 was corroborated with 95% confidence.

By applying Communication Analysis leads to functional requirements specifications with a more appropriate granularity than applying Use Cases (H41 was corroborated). This outcome may be explained by the fact that Communication Analysis methodological guidelines for Communicative Event Diagrams are based on more objective and prescriptive criteria than those of Use Case Diagrams.

6. Validation Evaluation: Threats

This section discusses issues with the potential to threaten the validity of the experiment [5].

6.1. Conclusion validity

We verified that the subjects had a homogeneous background by means of a questionnaire, so there is no threat due to random heterogeneity of subjects, which could give rise to greater variability in measures. As a trade-off, homogeneity limits external validity.

The proposed metrics have been theoretically reasoned and intend to be quite objective. However, an empirical testing of the metrics is advisable, in order to ensure the reliability of measures; this is planned as future work. Also, we plan to have the RS reviewed by more expert reviewers and to perform agreement rounds; this will allow assessing the level of objectiveness of the metrics and the measuring procedure (inter-reviewer agreement).

6.2. Internal validity

Instrumentation is the effect caused by the instruments used in the experiment. For instance, the experiment relies on the training material and seminars, but they may not represent the methods appropriately.

There is a risk related to the selection of subjects. Letting students decide which group to join according to their availability was a mistake; it led to imbalanced groups and the resulting groups cannot be assumed to have the very same characteristics (i.e. motivation).

We acknowledge a threat of maturation; that is, during the second round, the subjects already know the Photography agency problem statement. We intended to compare the results of the first round, but this was not possible due to the imbalance between both groups.

6.3. Construct validity

Two experimenters are authors of Communication Analysis. In order to reduce the threat of bias, two experimenters without expectancies have been involved.

A reference model of low quality is a threat. However, the authors have been using the Photography Agency case for research and education for more than ten years, and its conceptual model is highly agreed by now. A three-person expert modelling committee made the final adjustments.

The experiment includes a single IS description so it may under-represent the construct of all ISs. Since the subjects were trained in both methods, the results of the second round may be affected by their previous knowledge. Again, the imbalance between groups did not allow a comparison of the first round.

### Table 4. Paired Samples Test for errFra measure

<table>
<thead>
<tr>
<th>95% confidence interval of the difference</th>
<th>Sig. 2-tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>errFra_CA – errFra_UC</td>
<td>-0.92842</td>
</tr>
</tbody>
</table>

### Table 5. Wilcoxon signed-rank test for errAgg measure

<table>
<thead>
<tr>
<th>Items</th>
<th>N</th>
<th>Mean rank</th>
<th>Sum of ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>errAgg_CA – errAgg_UC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative ranks</td>
<td>20</td>
<td>13.63</td>
<td>272.50</td>
</tr>
<tr>
<td>Positive ranks</td>
<td>5</td>
<td>10.50</td>
<td>52.50</td>
</tr>
<tr>
<td>Ties</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) errAgg_CA < errAgg_UC, b) errAgg_CA > errAgg_UC c) errAgg_CA=errAgg_UC.

### Table 6. Test statistic-significance

<table>
<thead>
<tr>
<th>Z</th>
<th>Asymp. Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>errAgg_CA – errAgg_UC</td>
<td></td>
</tr>
<tr>
<td>-3.062</td>
<td>0.002</td>
</tr>
</tbody>
</table>

a) Based on positive ranks
6.4. External validity

With respect to the use of students as experimental subjects, several works suggest that, to a great extent, the results can be generalised to industry practitioners [42]. In any case, we are aware that more experiments with a larger number of subjects are necessary.

We thoughtfully selected a representative problem statement. However, more empirical studies with other requirements specifications are necessary.

7. Conclusions and Further work

Empirical evaluation of requirements specification (RS) techniques is a strong need in the area of Requirements Engineering. This paper extends a framework of conceptual model quality [9] by refining quality goals and proposing metrics that allow their operationalisation. The focus is put on semantic completeness and granularity. With regard to semantic completeness, most previous works propose a rating based on judgement (e.g. using a Likert scale). We propose measuring the degree of functional encapsulations completeness (degFEC) and the degree of linked communications completeness (degLCC) with respect to a reference model (made by an expert modelling committee). Also, we propose assessing whether a functional RS has an appropriate granularity with respect to a given unity criteria; this allows determining the number of functional fragmentation errors (errFra) and functional aggregation errors (errAgg).

We acknowledge that the proposed quality goals and metrics related to feasible functional completeness are more practical in experimental settings than in industrial practice (a reference model of the domain is very unlikely to be available). In any case, they provide a valuable strategy to feedback method designers.

A laboratory experiment has been carried out to compare two RE methods; namely Use Cases [6] and Communication Analysis [7]. Functional requirements specifications have been evaluated with respect to the proposed metrics. The following hypotheses were verified: Communication Analysis allows obtaining RS with greater degree of functional encapsulations completeness and greater degree of linked communications completeness than Use Cases; also, Communication Analysis allows obtaining RS with less functional aggregation errors than Use Cases. As regards functional fragmentation errors, the difference, although favourable for Communication Analysis, was not statistically significant. These outcomes can be due to the fact that Communication Analysis offers prescriptive methodological guidance to functional requirements encapsulation. We believe that Use Case-based methods would benefit from taking into account the underlying unity criteria [14].

Empirical evaluations (e.g. this experiment) allow comparing methods and highlighting their strengths and weaknesses. However, theoretical evaluations allow understanding better why these differences arise. This is planned for future research.

Future work also includes replicating this experiment, minimising the identified threats to validity, and validating the proposed metrics. We consider the possibility of basing future research in the extended framework by Krohgstie et al. [17]. This may facilitate fitting granularity in the framework. We also plan to present a case study describing the application of Communication Analysis to the integration of Anecoop S.Coop (a Spanish major distributor of fruit and vegetables) with its more than 100 associated cooperatives.

Acknowledgements

Research supported by the Spanish Ministry of Science and Innovation (MICINN) project SESAMO (TIN2007-62894), the MICINN FPU grant (AP2006-02323), and co-financed by FEDER.

We are also grateful to Dr. Lefteris Angelis from Aristotle University of Thessaloniki (Greece) for his helpful advice on statistical analysis.

References


