Language Quality in Information Systems Development – Analyzing the Emergence of Requirements in Natural Language Processes

Marianne Corvera Charaf
Goethe University Frankfurt, charaf@wiwi.uni-frankfurt.de

Follow this and additional works at: http://aisel.aisnet.org/amcis2010

Recommended Citation
http://aisel.aisnet.org/amcis2010/391

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2010 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Language Quality in Information Systems Development – Analyzing the Emergence of Requirements in Natural Language Processes

Marianne Corvera Charaf
Goethe University Frankfurt
charaf@wiwi.uni-frankfurt.de

ABSTRACT
A central challenge in information systems development (ISD) is the question how meaningful and coherent requirements can be developed as a consensual result of the interaction between analysts and clients. As language-based communication is the main venue for the system’s conceptualization, we assume that the understanding of how requirements emerge on the micro level of interaction is of critical importance. We claim that the quality of the communication between the different stakeholders depends on the ability to manage the question of how people deal with language in practice and reach consensus in the concrete ISD process. Therefore, in this paper we adapt the concept of quality to the question of how a common language is built in the ISD process. After defining a set of reasonable language quality dimensions, we apply our results to concrete observation cases from an ISD project and obtain insight beneficial to researchers and practitioners alike.

Keywords

INTRODUCTION
One of the most challenging aspects in the field of information systems development (ISD) is the fact that the product to be developed, especially in the requirements elicitation phase, exhibits a mostly intangible nature (Cule et al. 2000). In fact, the product can be described as the result of the interaction between the two groups of people usually involved in specifying, developing, managing, and controlling the requirements in ISD projects: (1) the users that make use of IT and (2) the developers that provide IT. Therefore, the question of how to overcome the so-called communication gap between these two groups is a critical success factor of ISD (Peppard and Ward 1998). Different cultures, different communities, and different languages confront the stakeholders with many uncertainties and ambiguities which influence the result, increase the costs and, in some cases, lead to failure of the whole process (Cule et al. 2000). Over many decades, these challenges have been addressed by IS researchers from different fields (e.g. Hirschheim et al. 1995; Lewis 1994; Sambamurthy and Kirsch 2000; Xia and Lee 2005). Nevertheless, results of various surveys (e.g. Agrawal and Chari 2007; Molokken and Jorgensen 2003) indicate that the challenge of building well-conceived ISs in a cost-effective way still exists today (Boehm and Basili 2000).

Rather than examining technical issues, we set the focus on the interaction between users and analysts, addressing the difficulties both have in establishing mutual understanding and communication in both directions (Alter 2001; DeMarco and Lister 1987; Hirschheim et al. 1995). As successfully creating a shared understanding between the stakeholders is a major driver for ISD project success (e.g. Gallivan and Keil 2003; Tan 1994), a central position in IS research is the understanding of IS as a process in the course of which a coherent and meaningful model is created by the consolidation of the different stakeholders’ perspectives and multiple requirements in an organization (Alvarez and Urla 2002).

In this research, we address this position by assuming that ISD is grounded in processes of language development and language formalization (Lyytinen 1985). However, we argue that the understanding of the processes on the micro level of interaction and communication can provide unique insights into the emergence of coherent and meaningful requirement specifications. At the core of this research lies the analysis of (1) how people use language in the ISD process and (2) by which language-related actions they achieve language quality.
Therefore we combine different perspectives and theories on language, communication and interaction, focusing on how language in an actual ISD project is shaped and regulated. Our contribution can be seen as a theory that helps to explain and predict the impact of linguistic actions in ISD (Gregor 2006).

In this research we firstly depict our fundamental theoretical propositions in the ISD field and explain how the concept of language quality is addressed. Secondly, after an overview of the research method and setting, we apply the language quality dimensions to concrete observation cases from an ISD project. After the presentation and discussion of results we conclude this work with an acknowledgment of limitations and suggestions for further research.

RELATED WORK

Adjacent to the functional tradition in IS research, since the 1980s new alternative approaches have considered the contribution of those involved in and affected by a system (Lucas 1975; Robey and Markus 1984) as well as the existence of a practice “fraught with volatility, exceptions, unstructured data and unpredictable requirements” (Truex et al. 2000). This led to a focus on communication-related questions about the processes of social interaction in IS and ISD (Hirschheim et al. 1995). Therefore, the analysis of methods to develop meaningful and coherent IS models implies that the resulting models have to be successfully legitimized on a language and knowledge level (Boland 1979) and consolidated by social interaction and communication (Hirschheim et al. 1991). This is supposed to happen throughout in a collaborative setting involving multiple stakeholders who represent the different fields of the organization as well as the system development section (Kavakli and Loucopoulos 2003).

At the core of these propositions we set the role of language as a formalization goal and a communication medium. In 1985, Lyytinen conceptualized the ISD process as a language development and formalization process (Lyytinen 1985). Not since then, this concept has been adopted by many IS researchers aiming to provide more insight into the development of meaningful, correct and sound IS models (Hansen and Rennecker 2006; Niehaves 2007; Weigand and Dignum 1997). An important position here is that to build well-defined semantic and syntactic model statements it is necessary to define their model element structure and terminological structure (Pfeiffer and Niehaves 2005). In this work we address the the semantical terminological structure, that is, the structure which gives the semantic relationship between symbols and concepts of the domain language (Becker et al. 2008). The domain language provides symbols and concepts relating to the particular domain of the real world to be represented. In IS literature, the relationship between the domain of the real world and the domain language is referred to as the “fragean core view” (Lyytinen 1985). According to this view, symbols of language have the function of denoting entities of the real world. Thereby a language based IS model or specification has the capacity to represent the domain of interest and is able to be the basis for ISD.

Furthermore, to achieve consensus on these model or requirements, stakeholders have to communicate and interact. Referring communication, we argue that natural language is the most probable common communication medium between stakeholders with heterogeneous backgrounds and the most frequently used communication instrument in practice (Janson and Woo 1995), even if the requirements are fixed later on in a formal language (Pohl 2007). Several contributions have been made discussing the role of natural language in ISD (Rupp and Sophisten 2002; Ryan 1993). The advantages of natural language are its universal use in different knowledge areas, its flexibility regarding abstraction grades and its simplicity of use (Kamsties 2001; Pohl 2007) whereas the disadvantages underline its inherent lexical, syntactical or semantic ambiguity (Pohl 2007). Allowing space for different interpretations of the same requirements may lead to an increase in negotiation costs and the risk of misunderstanding.

However, to analyze the process of language use, it is necessary to adapt interaction theories. Only a small number of researchers have focused on language use in the ISD interaction (e.g. Alvarez and Urla 2002; Hansen and Rennecker 2006; Marakas and Elam 1998). This view on language corresponds to the “ordinary speaking view” and rests on the assumption that the use of natural language can be observed as a social act, in which the stakeholders mediate their intentions (Lyytinen 1985). In the analysis of interaction related to language in the field of IS a major focus has been set on the Language Action Perspective (LAP) (Goldkuhl and Lyytinen 1982; Winograd 1988) which focuses on linguistic communication as the basis for understanding of ISs and on the impact of language action on a system. Nevertheless, the point where we set a different focus than LAP is the question whether language use is a precondition for or a result of the interaction. LAP has largely been viewing language as a precondition. In contrast, in this research we try to analyze the process of building a language with a view to language quality.
LANGUAGE QUALITY

While there are no doubts about the importance of communication between the different stakeholders in ISD, the issue of overcoming the ambiguity of natural language is frequently taken for granted.

We address this issue by applying the construct of a sign from the field of linguistics. A linguistic sign is understood as a combination of a concept – the signified (signifié) – and a sound image or symbol – the signifier (signifiant) (de Saussure 1974). According to de Saussure, a language consisting of linguistic signs is based on conventions related to the concept-symbol relationship as a precondition for meaningful language-based communication. In everyday language, these signs are used and actualized in changing combination and variation. Based on Language Critique (Kamlah and Lorenzen 1984), two main insights characterize the function languages have for communication in organizations. Firstly, in order to align meanings of symbols for domain specific communication, language constructs need to be introduced and explained. Second, the achievement of shared understanding makes a communication process more effective. These concepts from philosophy of language were applied to the IS field by Holten (2007) and Holten and Rosenkranz (2008), assuming that the immanent ambiguity of natural language is reduced by the joint construction of a “shared understanding” and a mutual basis of language.

In its simplest form this process comprises the constitution of a semantic relationship between a symbol (or a word) and its concept or description (Kamlah and Lorenzen 1984; Ortner 2005). Nevertheless, the idea that this process requires only one constituting action denies the fact that the venue in which meaning is created is the pragmatic level of language. In the words of Bühler, language is shaped “empractically”, which means, embedded in the interaction. On the one side this means that we usually do not have definitions completed by a single constituting action, that is, the complexity of a concept may require different explanation rounds to become consensus (Hoppenbrouwers and Weigand 2000). On the other side, the course of consensus building depends on the concrete interaction situation, where the different knowledge bases of the stakeholders influence the process. These implications make clear the complexity of achieving concise relationships between a symbol and a concept during conversations, which is our focus on quality.

To address the quality claim in the IS field, we adapt the data quality dimensions proposed by Wand and Wang (1996). Their definition of quality is dependent on the relationship between the relevant states of the real world and the states implemented in a system. For the field of language development in ISD we claim that the success of the implementation depends on the question how a relationship between the representing part and the meaningful part of language is built, that is: the ability of stakeholders to build a concise and meaningful relationship between single concepts and symbols involved in the ISD process during the communication process. As, according to Wand and Wang (1996), parts of the real world are described as states with limited values or attributes, in the same manner their language representations can be translated as concept-symbol relationships, which also have limited values or attributes. In the ISD process these concept-symbol relationships are shaped and this is the process in which, in the end, the claim of quality has to be fulfilled. Along with the concept of quality, Wand and Wang (1996) propose four data quality dimensions to be considered. Table 1 shows the adaptation of these quality dimensions to the concept of language quality.

<table>
<thead>
<tr>
<th>Language Quality Dimension</th>
<th>Description</th>
<th>Repair Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td>Completeness is given when for every concept a representing symbol is given.</td>
<td><img src="image" alt="Symbol" /> ➔ <img src="image" alt="Symbol" /></td>
</tr>
<tr>
<td>Meaningfulness</td>
<td>A meaningful representation is given when every representing symbol is linked with one corresponding concept.</td>
<td><img src="image" alt="Symbol" /> ➔ <img src="image" alt="Symbol" /></td>
</tr>
<tr>
<td>Non-redundancy</td>
<td>Non-redundancy is given when in every linguistic unit single concepts are linked to one concrete symbol.</td>
<td><img src="image" alt="Symbol" /> ➔ <img src="image" alt="Symbol" /></td>
</tr>
<tr>
<td>Unambiguosness</td>
<td>Unambiguosness is given when no two concepts map into the same symbol.</td>
<td><img src="image" alt="Symbol" /> ➔ <img src="image" alt="Symbol" /></td>
</tr>
</tbody>
</table>

Table 1. Language Quality Dimensions

Legend: concept symbol

OBSERVATION CASE AND RESEARCH METHOD

To better understand the emergence of language quality and to be able to evaluate the impact of actions on the different language quality dimensions, we analyzed the face to face communication in a concrete ISD project which aimed for the development of an application for analysis, storage and retrieval of market-specific and user-customized information. The project ran from December 2006 to September 2009 in Germany. The common language and the project language were German. The stakeholders involved were three participants from a project management team, six from an enterprise team, five from the developer team, and two from the requirements engineering team. The latter was the research team, which had a project internal role. The research was performed from December 2006 until January 2009. The meetings took place in different project member configurations.

The central data collection method was qualitative fieldwork. The analysis of the stakeholders’ language development interaction focuses on the analysis of the transcriptions and field notes of utterances from the meetings.

In the course of the research, we were able to cover nine meetings. All of them were observed by two researchers. Five meetings were audio-recorded and transcribed. From this data, we chose 37 observation cases (OC) with a language-defining character, that is, sequences in which project-relevant concept-symbol relationships were introduced or adjusted. For every OC we traced the language-defining sequences, which are sequences in which the common language basis was adjusted. Then, in the first coding round, for every OC we coded the data differentiating between the aforementioned constructs “symbol” and “concept”. In a second step, we then coded differentiating between the language quality dimensions.

RESULTS AND DISCUSSION

Completeness in Language Definition

In ten OCs, we found evidence for repair actions in the quality dimension of Completeness. The OCs usually began with a description of a concept, giving examples or specifying the functionalities. In OC-13 and OC-32, the descriptions of the concepts were followed by explicit requests for symbols to address the concepts. Consequently, corresponding symbols were defined.

In cases where no explicit requests for symbols were uttered, we identified two reactions. In OC-12 and OC-24, although no requests were uttered, corresponding symbols were spontaneously included in the feedback. For example, in OC-12 the question “Is it possible to make [description of the concept]?” elicited the response, “‘Automatic Tagging’ is too difficult”.

A different course of action was observed in OC-16, OC-25, OC-31 and OC-33, in which stakeholders also described new concepts without explicitly requesting symbols to address them. In these cases, the following interactions did not yield new symbols for the functions; instead, stakeholders reacted only with general feedback like an affirmative utterance. This raises the question, how self-evident is the endeavor to achieve completeness.

Although such a general feedback seemed to be sufficient sometimes, there were other cases where it led to insecurity. For example, in OC-16 we observed a time-consuming discussion breaking out because stakeholder 1 (SH 1) was not sure they were talking about the same concept:

SH 1: Is there something like in Wikipedia, where in the middle of the text you have blue text for relevant words and then you can click on it and it automatically leads you to another module [...]?
SH 2: Yes.
SH 1: I mean so that you can jump immediately to another topic...
SH 3: I do not know if that happens automatically.
SH 1: No, I mean...

In still other cases, like OC-34 and OC-35, the other stakeholders referred to new concepts by using attributes from the first stakeholder’s description, e.g.: “The ‘display thing’ is better” (OC-35). This seemed sufficient at the time to make clear what was meant, as all the stakeholders were involved in the ongoing discussion. Nevertheless it can be assumed that a couple of weeks later it may be difficult to refer back to the same concept as “the display thing” as this symbol is of a volatile nature.
From these observations we conclude:

- Speakers frequently do not pursue Completeness, there is no automatic mechanism prompting them to find a new symbol each time a new concept arises.
- The absence of a symbol can lead to insecurity in an ongoing interaction.
- The use of volatile referencing symbols can lead to future difficulties when referring back to the same concept at a later stage.

**Meaningfulness in Language Definition**

Repair actions regarding the dimension of Meaningfulness were identified in 13 cases. In OC-20, OC-21 and OC-23, stakeholders gave definitions of symbols spontaneously without other requesting them, e.g. “This is a 'Double-Log-In', that means [...]”. This allowed them to integrate concise concept-symbol relationships as new project requirements. In OC-9, OC-11, OC-17, OC-18, OC-28 and OC-30, we identified explicit questions for the meaning of symbols already given, e.g. “What do you mean by 'User'?”. This also made a discussion and acceptance of meaningful requirements possible.

Nevertheless, in OC-28 and OC-21, the symbols that were given seemed to appear so alien to some stakeholders that they had difficulties adopting them. In OC-28 for example the given symbol was “scale of volatility”. Later on, one of the stakeholders referred to it as “the vola... whatsit...”. From this we conclude that the use of a more easily accessible vocabulary can be helpful.

Another remarkable case was OC-27, in which the concept of the symbol “M-N-Relationship” was not understood by one of the stakeholders. He made it implicitly clear that there was a misunderstanding but did not explicitly articulate the reason for his confusion. After some rounds of discussion, he finally asked explicitly and clearly for a definition of the symbol “M-N-Relationship” (underlined sentence in the citation):

\[ SH 1: \text{Exactly that was the question of the M-N relationship between 'articles' and 'modules'. That is why I did it like that. Because it was... according to the ERM it was so, that is why I did it like that. } \]

\[ SH 2: \text{Well...} \]

\[ SH 1: \text{Is it clear what I mean?} \]

\[ SH 2: \text{When I take articles and... no... I don’t know...} \]

\[ SH 1: \text{That is the document, it is one article.} \]

\[ SH 2: \text{If I choose one article, does it show automatically the top the module or something like that? That is what I mean.} \]

\[ SH 1: \text{Yes, exactly.} \]

\[ SH 2: \text{Exactly...} \]

\[ SH 1: \text{That’s why, because it is a M-N relationship, it is not possible to deduce from one of them to the other.} \]

\[ SH 2: \text{Well, this M-N relationship... I don’t understand what it means...} \]

\[ SH 1: \text{Eh, different modules can be assigned to different articles. That is, different modules can be assigned to one article.} \]

Hence, our conclusions regarding the dimension of Meaningfulness are the following:

- If the concept denoted by a new symbol is not self-evident, an explicit request for its definition is necessary to conduct a successful interaction.
- Sometimes meaningful concept-symbol relationships may be abolished because some symbols are too difficult to adopt.
- Using an accessible vocabulary can be helpful.
- A clear request for a definition can avoid time-consuming discussions.

**Non-redundancy in Language Definition**

Regarding the dimension of Non-redundancy, we found four instances of repair actions. In all of them we identified two symbols linked to the same concept (OC-4, 5, 14, 37). In OC-4 for example there were discrepancies between the use of the names “field report” and “case study”, which both referred to the same concept (a kind of document in the IS). Although all stakeholders seemed to link both symbols to the same concept, the use of each symbol apparently brought confusion, as the stakeholders showed insecurity, e.g. by beginning a sentence using one symbol and ending it with the other. It took several meetings until the use of “field report” was implicitly adopted by all stakeholders.
Similarly, in OC-5, the two symbols “content elements” and “information component” referred to the same concept, which seemed to be clear to everyone. Still, even with these, at one point one of the stakeholders explicitly requested the future use of “content elements” as he had different associations with “information component”.

In both remaining cases (OC-14 and OC-47), one of the redundant symbols was actually also in use to refer to another concept. By listening to the chain of argumentation, in both cases one of the stakeholders noticed the incorrect use of the symbol and explicitly corrected the others. In both cases the stakeholder explained the incorrectness as a result of mixing up words. Here we also conclude that using a more easily accessible vocabulary can be a solution.

These observations lead to following conclusions:

- Achieving Non-redundancy in language definition can be a long process if not made explicit.
- Redundancy causes insecurity.
- Redundancy can result from confusion regarding additional concept-symbol relationships.
- The use of accessible vocabularies can be helpful.

### Unambiguousness in Language Definition

Repair actions regarding Unambiguousness were observed in five cases. In OC-1, OC-2, and OC-29 stakeholders explicitly asked for the clarification of possible misunderstandings. In OC-29 for example, the request for clarification was made immediately after the unclear use of the symbol “synonym group”. By using an example, the confused listener tried to verify if his perception was correct. In contrast, in OC-1 and OC-2 the discussion following the misunderstanding went on for quite some time (in OC-1 about two hours, in OC-2 until the next meeting). Although some discrepancies were evident, it took a long time until one of the stakeholders explicitly indicated the possibility of talking about different concepts.

In OC-10 and OC-26 we could not identify points at which the misunderstandings were explicitly cleared up.

The following conclusions result for the quality dimension of Unambiguousness:

- The explicit indication of a misunderstanding can accelerate the clarification.
- Sometimes a misunderstanding is not explicitly clarified.

### Conclusions and Propositions

We addressed the research goals to analyze (1) how people use language in the ISD process and (2) by which language-related actions they achieve language quality in two ways.

Firstly, we adopted the theoretical concept of data quality. Secondly, we applied this concept to concrete cases in an ISD project and developed several conclusions for every language quality dimension. Table 2 provides an overview of all therefrom resulting propositions, classified along the language quality dimensions. The positive or negative impact on the practice of the pursuit of language quality results from the presented interpretation of the observations.
<table>
<thead>
<tr>
<th>P#</th>
<th>Proposition</th>
<th>Dimensions</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Explicitness in the definition of symbols or concepts has a positive impact on time and is necessary for successful language interactions.</td>
<td>Meaningfulness, Non-redundancy, Unambiguousness</td>
<td>+</td>
</tr>
<tr>
<td>P2</td>
<td>A clear request for a definition of symbols or concepts can avoid time-consuming discussions.</td>
<td>Meaningfulness</td>
<td>+</td>
</tr>
<tr>
<td>P3</td>
<td>The use of accessible documents can be helpful.</td>
<td>Meaningfulness, Non-redundancy</td>
<td>+</td>
</tr>
<tr>
<td>P4</td>
<td>Striving for completeness in language seems to be not self-evident.</td>
<td>Completeness</td>
<td>-</td>
</tr>
<tr>
<td>P5</td>
<td>The lack of symbols to refer to concepts can lead to insecurity in ongoing interactions.</td>
<td>Completeness, Non-redundancy</td>
<td>-</td>
</tr>
<tr>
<td>P6</td>
<td>The use of volatile referencing symbols can lead to future difficulties when referring to the same concepts again in future.</td>
<td>Completeness, Non-redundancy</td>
<td>-</td>
</tr>
<tr>
<td>P7</td>
<td>Sometimes, meaningful concept-symbol relationships are abandoned because the symbols appear too complex to be easily adopted.</td>
<td>Meaningfulness</td>
<td>-</td>
</tr>
<tr>
<td>P8</td>
<td>The achievement of non-redundancy in language definition without any intervention can be a long process.</td>
<td>Non-redundancy</td>
<td>-</td>
</tr>
<tr>
<td>P9</td>
<td>Redundancy can result from confusion with other concept-symbol relationships.</td>
<td>Non-redundancy</td>
<td>-</td>
</tr>
<tr>
<td>P10</td>
<td>Sometimes misunderstandings are not explicitly clarified.</td>
<td>Unambiguousness</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Summary of Propositions for each Language Quality Dimension

LIMITATIONS AND OUTLOOK

At the core of this research we approached the questions of how people use language and how language quality can be assessed in the ISD process by adopting the concept of data quality (Wand and Wang 1996). By applying the developed language quality dimensions to a concrete ISD project, we were able to develop practical propositions which have a positive or negative impact on language quality.

For researchers, the application of the concept of language quality to the field of ISD appears to be a suitable method to analyze the emergence of concise and meaningful requirements. On the other hand, for practitioners it can be helpful to know which consequences the different actions undertaken in this field might have on language quality. While we cannot generalize the importance of language quality for other fields in which language plays a secondary role, we feel safe to assume that in our field, the issue has a relevance which needs to be assessed.

Although the observational data was limited to a single project and data analysis was restricted, the results of this research can provide qualitative evidence and have the capability to develop analytical and statistical generalization by the application to various and larger ISD projects. By adopting the concept of language quality we hope to provide a useful approach for future research in the IS field. Further research can also focus on the analysis of power relations or group dynamics in this field, which may have a negative or positive impact on language quality.

With this research we hope to provide a useful adaptation of language quality analysis and at the same time contribute to the practitioner’s view of the ISD language definition processes.
REFERENCES


