# Reducing Ambiguities in Requirements Specifications via Automatically Created Object-Oriented Models

Daniel Popescu[1] (dpopescu@usc.edu),
Spencer Rugaber[2] (spencer@cc.gatech.edu),
Nenad Medvidovic[1] (neno@usc.edu), and
Daniel M. Berry[3] (dberry@uwaterloo.ca)

[1] Computer Science Department, University of Southern California, Los Angeles, CA, USA
[2] College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[3] Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada

**Abstract.** In industry, reviews and inspections are the primary methods to identify ambiguities, inconsistencies, and under specifications in natural language (NL) software requirements specifications (SRSs). However, humans have difficulties identifying ambiguities and tend to overlook inconsistencies in a large NL SRS. This paper presents a three-step, semi-automatic method, supported by a prototype tool, for identifying inconsistencies and ambiguities in NL SRSs. The method combines the strengths of automation and human reasoning to overcome difficulties with reviews and inspections. First, the tool parses a NL SRS according to a constraining grammar. Second, from relationships exposed in the parse, the tool creates the classes, methods, variables, and associations of an object-oriented analysis model of the specified system. Third, the model is diagrammed so that a human reviewer can use the model to detect ambiguities and inconsistencies. Since a human finds the problems, the tool has to have neither perfect recall nor perfect precision. The effectiveness of the approach is demonstrated by applying it and the tool to a widely published example NL SRS. A separate study evaluates the tool's domain-specific term detection.

## 1 Introduction

The typical industrial software specifier writes software requirements specifications (SRSs) in a natural language (NL). Even if a final SRS is written in a formal language, its first draft is usually written in a NL. A NL SRS enhances the communication between all the stakeholders. However, on the downside, often a NL SRS is imprecise and ambiguous [3].

Many an organization follows a three-step review process to assess the quality of a NL SRS and to identify ambiguities and other defects in the NL SRS [35]. First, assigned reviewers try to find defects in the document. Second, in a meeting of the reviewers, all found defects are collected and rated according to their severities. Third, the reviewed NL SRS and the collected defects are sent back to the authors for corrections.

In this process, the quality of a review of a document is dependent mainly upon how effectively each human reviewer is able to find ambiguities and other defects in the document. However, a human reviewer and even a group of them have difficulties

identifying all ambiguities, even when using aids such as checklists. A human reviewer might overlook some defects while reading a SRS, because he might assume that the first interpretation of the document that came to his mind is the intended interpretation, unaware of other possible understandings. In other words, he unconsciously disambiguates an ambiguous document [10].

When a NL SRS is large, some ambiguities might remain undetected, because ambiguities caused by interaction of distant parts of the NL SRS are difficult to detect. In any one review sessions, only an excerpt of the NL SRS can be reviewed. Any ambiguity in the reviewed excerpt that is caused by interaction with a part of the NL SRS outside the excerpt may not be detectable. Moreover, when a NL SRS is large, lack of time may prevent some parts of the NL SRS from ever being reviewed. Having a faster method for conducting reviews would permit larger chunks of the whole NL SRS to be reviewed at once. It would permit also faster reviews so that more reviews and, thus, greater coverage of the NL SRS would be possible in any duration.

A controlled, grammar-constrained NL [9, 8] helps to reduce ambiguities by constraining what can be said in the NL. One possible grammar rule eliminates passive voice and, therefore, ensures that the doer of an action is always known. However, a controlled NL can address only syntactic ambiguity. Semantic ambiguity is beyond its control.

Because a semantic model omits unnecessary details, it helps to reduce complexity, and it helps the user to visualize important aspects. Therefore, with the help of a semantic model, a human reviewer can validate larger system descriptions than otherwise. Moreover, the human reviewer can focus on conceptual correctness and does not need to worry about the consistent use of concepts and correct grammar usage. Therefore, if a semantic model can be created of the system specified by a NL SRS, a thorough review of the NL SRS becomes easier.

Of course, constructing a model bears the risks of introducing new defects and of the model's not representing the original NL SRS. Researchers have tried to mitigate these risks by using automatic approaches and NL processing (NLP) [32, 25, 12] techniques, ranging in complexity from simple lexical processing [e.g., 7, 41, 40], through syntactic processing [e.g., 26, 34], all the way through language understanding [e.g., 11]. A software tool can scan, search, browse, and tag huge text documents much faster than a human analyst can. Furthermore, a tool works rigorously, decreasing the risk of overlooked defects and unconscious disambiguation. Each of some of the approaches tries to build a model from a NL source and may even try to reason about the content, possibly with assistance from the human user.

However, no automatic NLP tool is perfect. For any NLP tool $T$, there are utterances in its NL for which $T$'s underlying formalism is not powerful enough to process the utterances correctly. Generally, the more complex the processing in $T$, the greater the chances for not processing an utterance correctly. The measures of correct processing are *recall* and *precision*, which are defined in Section 4.2. For now, recall is a measure of how much of what $T$ should find it does find, and precision is a measure of how much of what $T$ does find it should find.

Certainly, the recall and precision of $T$'s parser bound $T$'s quality. In addition, the recall and precision of $T$'s domain-specific term (DST) identification bound $T$'s quality.

The typical domain has its own terms, abbreviations, and vocabulary. Therefore, a tool must detect these to create a correct model. Each fully automated tool has difficulty to create a correct model because it relies on a semantic network that is based on a predefined domain dictionary. For many a domain, a domain dictionary does not exist. A semi-automated tool requires its human users to build a domain dictionary while writing the NL SRS. Clearly, not every requirements engineering (RE) process is so mature that in it, a domain dictionary is built. Even in a mature process, some domain terms might be forgotten, because the analysts assume that these terms are widely understood and recognized.

We have created an approach for helping a specification writer or reviewer identify ambiguities in a NL SRS, in which the approach tries to address all of the above mentioned problems. Hereinafter, the new approach is called *our approach* to distinguish it from other approaches. For our approach, we have built a prototype dowsing[1] tool, called *Dowser*. Dowser is based on one controlled NL. It can create from any NL SRS an object-oriented (OO) diagram, which can then be assessed by human reviewers.

In the first step, Dowser parses a NL SRS and extracts the classes, methods, and associations of a textual class model from the NL SRS. In the second step, Dowser diagrams the constructed textual class model. In the third step, a human reviewer can check the generated diagram for signs of defects in the NL SRS. Dowser and our approach are based on NL processing (NLP) and not on NL understanding. Dowser cannot judge whether or not the produced model describes a good set of requirements or classes. This judgement requires understanding. Therefore, in our approach, a human is the final arbiter of ambiguity. Because the human is in the loop, Dowser has to have neither perfect recall nor perfect precision.

To evaluate the effectiveness of our approach, we have implemented a prototype of Dowser and have tested it on a widely used example SRS, describing an elevator system [15]. The case study demonstrates the effectiveness of the approach. Since identifying domain terminology is required for any successful NLP-based approach, we conducted separate studies to evaluate Dowser's DST detection. The studies show that our approach is capable of achieving high recall and precision when detecting DSTs in a UNIX manual page.

Section 2 discusses related work. Section 3 describes out approach and all of its components. Section 4 describes validating case studies, and Section 5 concludes the paper by discussing the results and future work.

## 2 Related Work

Related work can be divided into four parts: (1) NLP on SRSs, (2) controlled languages, (3) automatic OO analysis model (OOAM) extraction, and (4) domain-specific term (DST) extraction.

**NLP on SRSs**: Kof describes a case study of the application of NLP to extract and classify terms and then to build a domain ontology [20]. This work is the most similar

[1] A dowser is a tool that makes use of domain knowledge in understanding software artifacts [5].

to our approach. The built domain ontology consists of nouns and verbs, which constitute the domain's concepts. In the end, the domain ontology helps to detect weaknesses in the requirements specification. Gervasi and Nuseibeh describe their experiences using lightweight formal methods for the partial validation of NL SRSs [12]. They check properties of models obtained by shallow parsing of natural language requirements. Furthermore, they demonstrate scalability of their approach with a NASA SRS.

**Controlled languages**: Fuchs and Schwitter developed Attempto Controlled English (ACE) [9], a sublanguage of English whose utterances can be unambiguously translated into first-order logic. Over the years, ACE has evolved into a mature controlled language, which is used mainly for reasoning about SRSs [8]. Juristo et al. developed other controlled languages, SUL and DUL [17]. For these languages, they defined a correspondence between linguistic patterns and conceptual patterns. After a SRS has been written in SUL and DUL, an OOAM can be created using the correspondence.

**OOAM Extraction**: Several tools exist that automatically transform a SRS into an OOAM. Mich's NL-OOPS [26] tool first transforms a parsed SRS into a semantic network. Afterwards, the tool derives an OOAM from the semantic network. Delisle, Barker, and Biskri implemented a tool that uses only syntactic extraction rules [6]. Harmain and Gaizauskas implemented another syntax-based tool, and they introduce a method to evaluate the performance of any such tool [15]. As in our approach, Nanduri and Rugaber [29] have used the same parser and had the same initial idea of using syntactic knowledge to transform a NL SRS into an OOAM. The objective of their approach was to validate a manually constructed OOAM. Our approach's main objective is to identify ambiguity, inconsistency, and underspecification in a NL SRS. The more restricted objectives of our approach enables a more detailed discussion of the problem space and contributes (1) a constraining grammar, (2) analysis interpretation guidelines, (3) additional transformation rules, and (4) DST extraction.

**DST extraction** Mollà et al. developed a method for answering questions in any specified technical domain. This work recognizes the importance of dealing with specified technical terminologies in NLP tools that are applied to SRSs [28, 11].

## 3 Our Approach

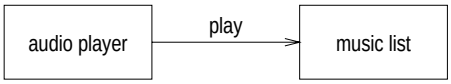
The goal of our approach is to reduce the number of ambiguities, inconsistencies, and underspecifications in a NL SRS through automation. Assuming that automation will not be perfect, i.e., it will have less than 100% recall and less than 100% precision, we let a human make the final decision about a potential ambiguity, inconsistency, or underspecification.

While reading through a NL SRS, an engineer usually builds a mental model of the described system and reasons about the correct relations of the SRS's concepts. If an engineer could only analyze the correctness of a model, instead of having also to create it, write it down, and then analyze it, he could use his skills and time more effectively.

Considering that a reviewer could more effectively inspect a model than the complete NL SRS, we developed an automatic approach based on the following observations:

- Each of most software companies uses a NL SRS [24] to describe a software system regardless of its domain.
- An OOAM is able to show the most important concepts and relations among the concepts of a system to build.
- Many an OO design method suggests building an OOAM of a sentence by identifying the parts of the sentence and creating a class from the subject, attributes from the adjectives, and methods and associations from the verb [1, 30].

Consider the example transformation of the sentence[2] The audio player shall play the music list. into an OOAM; audio player is the subject of the sentence, play is the verb, and music list is the direct object. This sentence could therefore be modeled as the diagram:




In this example, the adjectives audio and music are not broken out, because each is part of a DST.

Using this syntax-based method, the typical functional requirement sentence of a NL SRS can be transformed into an OOAM. Since this heuristic suggests using mostly syntactic information, the transformation can be automated. A NL parser can create parse trees of any NL text [34]. The OO design literature gives many rules and heuristics to transform many a syntactic construct into a textual OOAM. Off-the-shelf software exists to diagram textual OOAMs [36].

Since NL SRSs are written for a wide range of domains such as medical, technical or judicial domains, a successful approach must be robust in identifying DSTs. Our approach addresses this need by using syntactic information and a robust parser with guessing capability.

The overall quality of any NL SRS can be improved by enforcing the use of a constraining grammar. A constraining grammar reduces the possibilities of ambiguities

[2] A sans serif typeface is used for example text, except in a parse tree. Beware of punctuation, also typeset in the sans serif typeface, at the end of any example. It should not be considered as punctuation in the containing sentence, which is typeset in the serifed typeface. Sometimes, two consecutive punctuation symbols appear; the first, typeset in the sans serif typeface ends an example, and the second, typeset in the serifed typeface is part of the sentence containing the example. A typewriter typeface is used for example text in any parse tree, in which monospacing is essential for the correct display of the tree.

by constraining the allowed language constructs. At the same time, it increases the quality of parsing, reduces the number of parses, and results in more precise OOAMs.

Therefore, by using and extending existing technology, we can create a tool that automatically transforms a NL SRS into an OOAM that helps a human being to identify ambiguities, inconsistencies, and under specifications in the NL SRS.

Figure 1 shows the flow of our approach. First, Dowser parses a NL SRS according

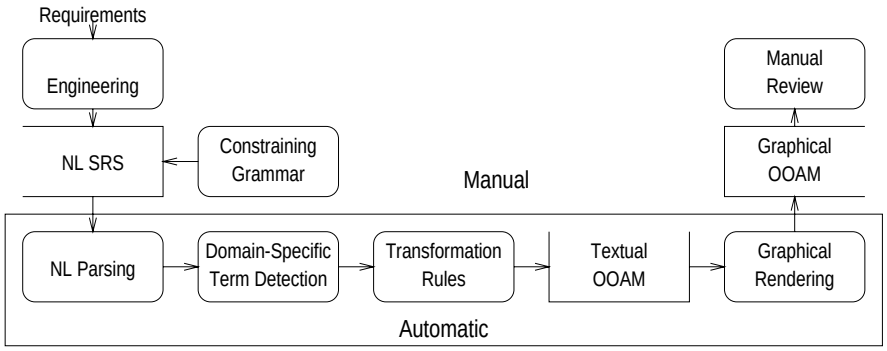



**Fig. 1.** Flow of the Approach

to a constraining grammar. Second, from relationships exposed in the parse, Dowser creates the classes, methods, variables, and associations of an OOAM of the specified system. Third, the OOAM is diagrammed so that a human reviewer can use the model to detect ambiguities, inconsistencies, and underspecifications.

One issue that arises from basing the model building, a semantic process, on a parser, a syntactic process, is "Why not have its parser report also syntactice ambiguity?", especially since whatever parser is used probably can report all parses it finds for any sentence. In the end, an integrated tool might very well report both syntactic and semantic ambiguities. However, given the wealth of work on syntactic ambiguity [e.g., 4, 7, 18, 25, 41] and the dearth of work on semantic ambiguity, the focus of this work is on semantic ambiguity. If and when a completely satisfactory approach to identifying semantic ambiguity is found, if it is based on a parser, then certainly that same parser can be used as the basis for identifying syntactic ambiguity in a single integrated ambiguity identification tool.

### 3.1 Constraining Grammar

Any NL allows expressing the same concept in different ways using different syntactic structures. For example, a sentence in active voice can be translated into passive voice without changing the semantics or pragmatics. However, passive voice can encourage ambiguities. For example, the sentence The illumination of the button is activated.

leaves room for different interpretations, because it is not clear who holds the responsibility for activating the illumination. Alternatively, the sentence could be describing a state. As a consequence, a constraining grammar can be introduced to decrease the possibility of ambiguity. A constraining grammar enables formal reasoning without the disadvantages of a fully formal language [8]. A constraining grammar has the other advantage that it is more amenable to parsing, and extraction rules based on it can be created more easily.

Observe that the constraining grammar used may bear no relationship with the internal grammar of the parser used other than sharing the same NL. The goals of the two grammars are different. A constraining grammar tries to reduce the number of ways to say something and to be uniguous[3]. The goal of a NL parser's grammar is to be as general as possible and to recognize any legitimate sentence in the NL. That is, a NL parser's grammar is designed to be as ambiguous as is the NL itself.

Our approach uses a constraining grammar that is derived from Juristo et al.'s grammar [17]. They have developed two context-free grammars and an unambiguous mapping from these grammars to OOAMs. This mapping is explicitly defined and allows better model creation than with commonly used heuristics that are justified only intuitively. Moreover, the explicit definition enables automation.

Using a constraining grammar influences the style of a NL SRS, because a constraint grammar enforces simple sentences. The typical sentence has a basic structure consisting of subject, verb and object. Furthermore, only simple subclause constructions are allowed, such as conditional clauses, using if, when, velc.[4] Therefore, a NL SRS will contain many short sentences if it is written according to the developed controlled grammar. Shorter, simpler sentences tend to be less ambiguous, because at the very least, they avoid some coordination and scope ambiguities

### 3.2 Natural Language Parsing

Since an OOAM is created automatically from syntactic information, we needed a parser to extract this information from the NL SRS. The parser we used was developed by Sleator and Temperley (S&T) at Carnegie-Mellon University [34]. Sutcliffe and McElligott showed that the S&T parser is robust and accurate for parsing software manuals [38]. Since software manuals are similar to SRSs [2], the S&T parser looked promising for our approach. Additionally, the S&T parser was chosen because it is able to guess the grammatical role of unknown words. This capability is used for the DST detection, which is described in Section 3.4. However, in principle, any other parser e.g., The Stanford Parser [19] could be used. Dowser would have to be adjusted to work with the parser's output.

The parser is based on the theory of link grammars, which define easy-to-understand rule-based grammar systems. A link grammar consists of a set of words, i.e., the terminal symbols of the grammar, each of which has one or more linking requirements. A sequence of words is a sentence of the language defined by the grammar if there exists a way to assign to the words some links that satisfy the following three conditions:

[3] The coined term "uniguous" means "not ambiguous".

[4] "velc." means "or others" and is to "vel cetera" as "etc." is to "et cetera".

1. Planarity: the links do not cross;
2. Connectivity: the links suffice to connect all the words of the sequence together; and
3. Satisfaction: the links satisfy the linking requirements of each word in the sequence.

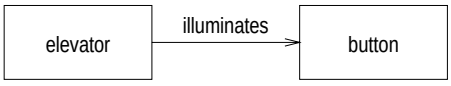
The link grammar parser produces the links for every such sentence. After parsing, the links can be accessed through the API of the link grammar parser.

Each established link in a sentence has a link type, which defines the grammatical usage of the word at the source of the link. The sentence The elevator illuminates the button. shows three different link types:

```
     +-------Os------+
 +---Ds--+-----Ss----+        +--Ds--+
 |       |           |        |      |
 the elevator.n illuminates.v the button.n .
```

A `D` link connects a determiner to a noun, an `S` link connects a subject noun to its finite verb, and an `O` link connects a transitive verb to its object. A small `s` after the type of a link indicates that the target of the link is a singular noun.

From this example sentence, the first extraction rule can be derived. If a sentence contains an `S` link and an `O` link, then create a class from the subject noun and one from the object noun. Afterwards, create a directed association from the subject class to the object class, which is named by the verb:




A directed association was chosen over a simple class method, because a directed association shows also who invokes the action. If the action were modeled as a class method, the information about who causes the action would have been lost. Using this rule, Dowser would extract the classes elevator and button and a directed association illuminates from the elevator class to the button class.

To avoid having two different classes created for elevator and elevators, our approach incorporates the lexical reference system *WordNet* [27] to find the stems of nouns and verbs. Therefore, the name of each created class is the stem of a noun.

One might think that the use of a constrained language would reduce the number of parses and might even ensure that there is only one per sentence. However, the language constraints only encourage and do not guarantee uniguity. In general, the link parser returns multiple parses for any input sentence. For example, When an elevator has not to service any requests, the elevator remains at its final destination and the

doors of the elevator are closed. returns 16 parses. However, as the link grammar homepage [39] says, "If there is more than one satisfactory linkage, the parser orders them according to certain simple heuristics." In our experience, these heuristics selected as first our preferred parse. Of courser, these simple heuristics cannot be expected to always select the parse that the writer intended, because syntactic ambiguity is a hard problem. Even humans have to rely on semantic understanding and context to resolve difficult cases of syntactic ambiguity, e.g., the classic The boy saw the man with the telescope..

### 3.3 Transformation Rules

Transformation rules bridge the gap between the extracted syntactic sentence information and the targeted OOAM. Each transformation rule describes how a combination of words of identified grammatical roles can be transformed into classes, associations, attributes, and methods.

The transformation rules Dowser uses were derived from Juristo *et al.*'s grammar [17], the OO methods literature [29, 30], and conducted experiments. In total, Dowser uses 13 transformation rules. The five most frequently used are the following:

1. The most frequently applicable rule is the first extraction rule, described in Section 3.2. If a parsed sentence contains a subject and an object link, then create two classes with a directed association named after the verb.
2. Aggregations are an important notion in UML class diagrams. One rule for extracting aggregations is similar to the first rule. The major difference is the verb. This rule is applicable only if the parsed sentence contains a subject and an object link and the verb stem is one of have, possess, contain, or include. In that case, the object is aggregated to the subject.
3. Sometimes, a subclause describes a system action without the need of an object, particularly, if the system reacts to a given event, e.g., If the user presses the button, the elevator moves.. An event clause starts with an if or when. If Dowser detects an event clause, and the main clause has only a subject link, then a class from the subject link noun is created and the verb is added to the new class as a method.
4. A genitive attribute indicates two classes with an aggregation, e.g., The system stores the name of the customer. or The system stores the customer's name.. If Dowser detects a genitive, it creates two classes with one linking aggregation. For either example, Dowser would create a class customer, and it would aggregate the class name to the class customer; name could have been modeled as an attribute. However, other sentences in the specification would add methods to the class name later. Therefore, with the syntactic information of only one sentence, it cannot be decided if name is an attribute or an aggregated class. This rule needs to be constrained by semantic information. For example, Dowser should not apply this rule to the sentence The user enters the amount of money.. Although amount is a noun, the class amount is not desired in this case.
5. Although active clauses are preferred in NL SRSs, passive clauses are still needed. They are used to describe relations and states, e.g. as in A husband is married to

his wife.. From this sentence, two classes are created, from the subject noun and the noun of the prepositional phrase. The passive verb and the connecting word to link the prepositional phrase described with the association.

Dowser applies two post-processing rules after it executes all possible transformation rules.

The first post-processing rule converts all classes that are aggregated to another class into attributes of that other class. Only a class that lacks any attribute, method, or incoming or outgoing association is transformed. For example, one rule described above extracts two classes and one aggregation from the sentence The system stores the name of the customer.. The rule creates a class name and a class customer. However, the class name has probably no method or association. Therefore, if a class contains no method after all rules have been applied, it is transformed into an attribute of the class customer.

The second post-processing rule removes the class system from the OOAM, since all other classes together form the system; system is not a class, because it cannot be a subpart of itself.

The full set of rules, a user's manual for Dowser, and other details may be found at the Web site, http://www.cc.gatech.edu/projects/dowser/.

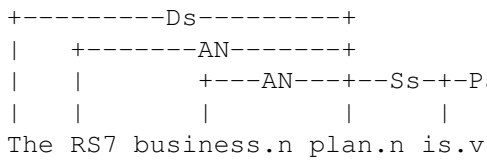
### 3.4 Domain-Specific Terms

In order to build the correct classes into the OOAM of a NL SRS, our approach has to be able to detect DSTs. If Dowser were to extract only the concept button from elevator button, Dowser would be identifying an incorrect term.

To achieve high DST recall, the parser could access a special domain data dictionary. However, for each of most domains, such a data dictionary does not exist. Software is built for financial, medical, technical, and other domains. Creating a domain dictionary for each of these rich domains is complex and difficult. Additionally, the customer of a software application might prefer to use her own terms for describing her product. A product could have arbitrarily chosen names, such as DNAComp07. Therefore, even if a domain dictionary exists for a NL SRS, DST detection remains a challenge.

The link types of the link grammar parse can be used to identify DSTs. The typical DST happens to be built from an attributive noun or a proper noun. Therefore, in a link grammar parse, the AN link, denoting an attributive noun, and the G link, denoting a proper noun, help to identify DSTs.

Consider the link grammar parse of the sentence The RS7 business plan is due in May.:



Thus, RS7 business plan is a domain-specific term.

A parser typically has problems parsing words that are not in its internal dictionary. However, the S&T link grammar parser has a guessing mode, in which it can guess the syntactic role of an unknown term. Therefore, it is often able to guess the syntactic role of an unknown term, improving its DST recall.

Since DST detection is essential for transforming a NL SRS into an OOAM, we conducted a study, described in Section 4.2, about the recall and precision of our approach.

### 3.5 Diagramming OOAMs

The previous steps are able to create a textual OOAM. However, it is easier for a human to understand a graphical OOAM than a textual OOAM. Using the approach described by Spinellis [36], an extracted textual OOAM is diagrammed. The tool *UMLGraph* [37] transforms a textual description into a *dot file*, which the tool *Graphviz* [13] can transform into any of some popular graphic formats, such as JPEG, GIF, or PNG.

### 3.6 Interpretation of OOAM

In the last step of our approach, a human analyst checks the created diagram for ambiguities.

Some ideas that a human analyst can use to find defects in an OOAM are:

- An association is a hint for possible ambiguities. For example, suppose that each of two different classes sends a message to the same target class. The analyst should check that the two different classes actually are to communicate with the same target class. If a motion sensor activates one type of display, and a smoke detector activates another type of display, then the class diagram should reflect this situation with two different display classes.
- Each class should reflect one and only one concept. For example, the analyst should check that book and textbook are really two different classes when Dowser does not create a generalization of these two classes.
- If a class has an attribute, but the attribute is not of a primitive type, such as string or number, then the definition of the attribute might be missing in the original text. After a definition is added, the attribute should be represented by its own class.
- If a class has no association, then the class might be underspecified, as there are no relations or interactions between the class and other classes.

### 3.7 Limitations of Method

Observe that the OOAM is a model of only static relationships among the concepts mentioned in the parsed NL SRS. We have not attempted to apply our approach to modeling behavior.

## 4 Studies

This section describes the case studies in which we evaluated the effectiveness of our approach in helping an analyst to identify ambiguities, inconsistencies, and underspecifications in a NL SRS and in which we evaluated Dowser's effectiveness at DST identification.

### 4.1 Elevator Case Study

To evaluate the effectiveness of our approach, we implemented Dowser and applied it to an example NL SRS that we call "the ESD". The ESD describes the control software for an elevator system [16]. The ESD was chosen, because it could be the NL SRS of a real industrial system. At the same time, the ESD is short enough to completely described in this paper. Moreover, the ESD happens to contain enough defects that it illustrates the defect types that can be revealed with the help of Dowser.

The original ESD was:

An n elevator system is to be installed in a building with m floors. The elevators and the control mechanism are supplied by a manufacturer The internal mechanisms of these are assumed (given) in this problem.

Design the logic to move elevators between floors in the building according to the following rules:

1. Each elevator has a set of buttons, one button for each floor. These illuminate when pressed and cause the elevator to visit the corresponding floor. The illumination is cancelled when the corresponding floor is visited (i.e., stopped at) by the elevator.
2. Each floor has two buttons (except ground and top), one to request an up-elevator and one to request a down-elevator. These buttons illuminate when pressed. The buttons are cancelled when an elevator visits the floor and is either travelling the desired direction, or visiting a floor with no requests outstanding. In the latter case, if both floor request buttons are illuminated, only one should be cancelled. The algorithm used to decide which to serve first should minimize the waiting time for both requests.
3. When an elevator has no requests to service, it should remain at its final destination with its doors closed and await further requests (or model a "holding" floor).
4. All requests for elevators from floors must be serviced eventually, with all floors given equal priority.
5. All requests for floors within elevators must be serviced eventually, with floors being serviced sequentially in the direction of travel.

First, we applied Dowser tool to the original unmodified ESD, which was not written according to any constraining grammar. Since the transformation rules have been created assuming that the analyzed text conforms to a constraining grammar, applying Dowser to the original ESD resulted in a diagram with only five classes such as these and set. None of these classes describes any domain concepts of the ESD.

To successfully apply Dowser to the ESD, the ESD had to be rewritten sentence-by-sentence to conform to the constraining grammar. No information was added or removed from the original ESD during the rewriting. Therefore, the rewriting did not introduce any new defects, which would have adulterated the results of the case study.

The rewritten ESD is:

An n elevator system is to be installed in a building with m floors.

1. Each elevator has buttons. Each elevator has one button for each floor. When a user presses a button, the elevator illuminates the button and the elevator visits the corresponding floor. When the elevator visits a floor, the elevator cancels the corresponding illumination.
2. Each floor has two buttons. (except ground and top). If the user presses the up-button, an up-elevator is requested. If the user presses the down-button, a down-elevator is requested. If the user presses a button, this button becomes illuminated. When an elevator visits a floor, the elevator cancels the corresponding illumination of the button in the desired direction. The system minimizes the waiting time.
3. When an elevator has not to service any requests, the elevator remains at its final destination and the doors of the elevator are closed. The elevator then awaits further requests.
4. The elevators service all requests from floors with equal priority eventually.
5. If a user presses a button within the elevator, the elevator services this request eventually in the direction of travel.

Applying Dowser to the new ESD resulted in the diagram of Figure 2. Dowser created an OOAM containing 14 partially connected classes with attributes and methods.

The graphically rendered OOAM does not reveal defects on its own. By applying the guidelines described in Section 3.6, we identified four conceptual defects in the OOAM, which could be traced back to the original ESD.

1. The diagram shows classes up-elevator and down-elevator. Neither class has any connection to any other class, and neither has any association to the class elevator. Furthermore, the up-elevator class has an attribute requested, while elevator serves a request indicates that each of up-elevator and down-elevator is a specialization of elevator. All of this information indicates that neither concept, up-elevator nor down-elevator, is defined enough in the original ESD.
2. The class door in the diagram contains the attribute closed. However, the class has no method to manipulate this state. If closing and opening the door are within the scope of the system, then it is clear that the concept door is not defined enough in the original ESD.
3. In the ESD, each of the floor and the elevator has buttons. Therefore, each of the class elevator and the class floor should have an aggregated class button in the OOAM. However, the diagram indicates that both have the same type of button. Since a button in an elevator and a button in a floor have different behaviors, it

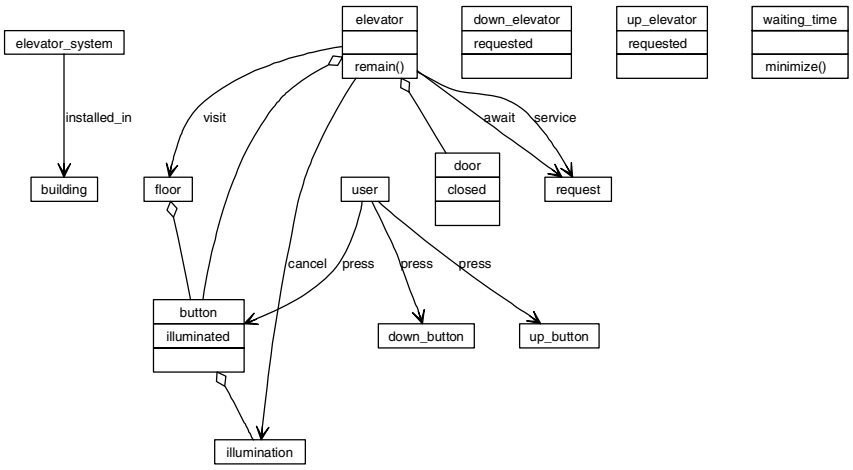



**Fig. 2.** OOAM of the ESD

is unlikely that the same class describes both types of buttons. Generalizing both types to a single class button could therefore lead to misinterpretations. Defining concepts elevator button and floor button would resolve this ambiguity and enhance the clarity of the ESD.

4. Each of the classes up-button and down-button is connected to only the class user in the OOAM. Since a user is an actor in the system, the diagram does not clarify where button belongs. The location can be derived from the ESD, because button is mentioned in the paragraph that mentions floor. However, it should not be necessary to use this fact. Therefore, each concept should be specified in more detail in the ESD to reduce the ambiguity in the specification.

This case study shows how Dowser can help an analyst identify defects in a NL SRS. If a constraining grammar is used to write a NL SRS, our approach can help detect ambiguities, inconsistencies, and underspecifications.

### 4.2 DST Detection Quality

The case study in Section 4.1 shows the importance of detecting DSTs. For the ESD, Dowser needed to be able to detect DSTs such as floor button, elevator button, or down elevator. If Dowser were to extract only the terms button and elevator from the ESD, it would create wrong classes and associations.

Section 3.4 explains how Dowser relies on syntactic information to identify DSTs. To measure Dowser's DST detection capability, we conducted a separate study. To measure Dowser's DST detection, the metrics *recall* and *precision* were calculated of Dowser's extraction of DSTs from a user's manual. These metrics are used to evaluate the success of many NLP tools [14] :

– *Recall* measures how well a tool identifies desired pieces of information in a source:

$$Recall = \frac{I_{correct}}{I_{total}} \tag{1}$$

Here $I_{correct}$ is the number of correctly identified desired pieces of information in the source text and $I_{total}$ is the total number of desired pieces of information in the source text.

– *Precision* measures how accurately a tool identifies desired pieces of information in a source:

$$Precision = \frac{I_{correct}}{I_{correct} + I_{incorrect}} \tag{2}$$

Here $I_{correct}$ is as for *Recall* and $I_{incorrect}$ is the number of incorrectly identified desired pieces of information in the source text.

We chose the *intro man page* of the *Cygwin environment* with which to measure recall and precision of Dowser's DST detection. A manual page seems to be a suitable experiment source, because it is a technical document with a large number of DSTs.

The steps of the experiments are as follows.

1. We manually identified every DST, a noun or a compound noun, from the *intro man page*. The *intro man page* contains 52 terms like Cygwin, Linux-like environment, Linux API emulation layer, and POSIX/SUSv2. The 52 terms consists of 33 single-noun terms and 19 compound-noun terms.
2. For the first experiment, the link grammar parser extracted every term out of the *intro man page* without the capability of extracting compound-noun terms. It recognized 31 of the single-noun terms and none of the compound-noun terms. Therefore, it reached a recall value of 59.6% for all terms and of 93.9% for single noun terms.
3. For the second experiment, compound-noun term detection was added to the link grammar parser. After this, the tool recognized 10 compound-noun terms. As the single-noun terms detection rate stayed the same, the tool recognized 41 terms. Therefore, it reached a recall value of 78.8%.

Afterwards, the undetected terms were examined. It turned out that five terms were undetected because they appeared in grammatically obscure or wrong parts in the sentence. Correcting these sentence, increased the detected terms to 46 and the recall value to 88.46%.

The five not identified terms were (1) case-insensitive file system; (2) intro man page; (3) Linux look and feel; (4) Red Hat, Inc.; and (5) User's Guide. The term Red Hat, Inc. is not recognized because of the comma, User's Guide cannot be detected syntactically, because if every genitive were part of a term, it would lead to an over-generation of terms. Linux look and feel is not recognized because of the conjunction and in the term. Case-insensitive file system and intro man page can be only partially detected, because case-insensitive is an adjective, which is only sometimes part of a term. Another example demonstrates the difficulties caused by adjectives. In readable manual, only manual is a term in most cases. Using every adjective as part of the term would lead to an overgeneration of terms. The term intro man page is not

recognized because the link grammar parser guesses that intro is an adjective. However, if it is planned that case-insensitive file system is a concept within a SRS, then writing it with initial upper-case letters would allow the link grammar parser to detect it as a proper noun, and thus as a DST.

Dowser extracted seven wrong terms, since it created wrong terminology for the incompletely detected terms, e.g., it extracted the term man page instead of intro man page. Overall, Dowser reached a precision value of 86.79% on the *intro man page*.

The DST detection experiment shows that using only syntactic information from the link grammar parser allows a fairly high DST detection rate.

### 4.3 Monterey Workshop Airport Security Case Study

As a second case study, we have applied Dowser to the workshop case study, here called WSCS, taken from the 2007 *Monterey Workshop on Innovations for Requirements Analysis: From Stakeholder Needs to Formal Designs* [22]. Unlike the case study reported in Section 4.1, the WSCS was not written as a SRS. The WSCS is a transcript of a discussion about airport security by three principal stakeholders: the Transportation Security Administration (TSA), the Federal Aviation Administration (FAA), and Airport Screening and Security (ASS). No stakeholder in the discussion even attempts to be correct, complete, and consistent. While no real SRS is correct, complete, and consistent, at least an SRS's authors try to make it so. One's goals in any writing effort affect the output.

The first step was rewriting the sentences of the WSCS discussion transcript to meet the constraints of the constraining grammar. The first main change was to make each sentence active [31]. The WSCS's domain is familiar enough to the man-in-the-street that we had no problems identifying each doer. In the normal industrial case, we would have asked the client for this information. The second main change was to remove each modal verb, such as can in can deal that consists of a modal verb followed by a main verb. A modal verb has no real effect on the OOAM to be built; the main verb carries all the meaning. There were some so-called sentences of the transcript that had to be made into more than one sentence. There were portions of the transcript that defied rewriting. They were simply left out.

The revised transcript had 65 sentences, and they were submitted to Dowser. The Dowser tool was able to parse and extract information out of 58 of these sentences, for an 89% acceptance rate, leaving 7 unaccepted sentences.

Since Dowser listed each unaccepted sentence, we were able to perform an additional detailed manual analysis. By creating the link-grammar parse of each sentence, we were able to compare the created links to the rules and found the reason for Dowser's rejection of the sentence. The reasons can be categorized into three types.

1. The link-grammar parser was not able to guess the correct syntactic role of a word in the sentence,
2. Dowser's rules were not refined enough to handle the sentence, or
3. the sentence was underspecified.

We rewrote each of the 7 sentences, keeping the semantics consistent with sentence's intent while clarifying underspecfications. We chose to rewrite the not accepted sentences,

because this strategy is the easiest way for an analyst to achieve a 100% acceptance rate with Dowser.

The final list of sentences from which Dowser can parse and extract information is given by the transcript below. Note that neither the first 5 lines nor any line with a three-letter acronym for a stakeholder followed by a colon were submitted to Dowser. These help the reader see the correspondence to the original transcript.

Case Study: Air Traveling Requirements Updated (Blog scenario)
Participants:
Transportation Security Administration (TSA)
Federal Aviation Administration (FAA)
Airport screening and security (ASS)

FAA:
Airline passengers not take liquids on board.
We increase security following the recent foiled UK terrorist plot.
We develop technologies that screen for chemicals in liquids.
You know backscatter.

ASS:
Technologies that work in laboratories cause false alarms when used dozens of times in daily screening.
Technologies are not ready for deployment.

FAA:
False positives help us to stay alive.
You be more alert.
ASS miss guns and knives packed in carry-on luggage.

ASS:
ASS moves 2 million passengers through US airports daily.
ASS not remain alert to rare events.

TSA:
TSA deal with it.
You take frequent breaks.
As a test, TSA will impose the image of a weapon on a normal bag.
ASS learns that the image of a weapon appear on a normal bag.
Solutions will be a combination of machine intelligence and human intelligence.

ASS:
ASS take breaks.
ASS get inspected.
ASS not get annual surprise tests.
ASS gets tests every day.
If a screener misses too many tests consistently, he is sent to training.

TAS:
TSA and ASS and FAA take proactive steps to prevent another attack.
We only react. We do not anticipate.
If someone uses a box cutter to hijack a plane, then passengers not take box cutters on board.
If someone hides explosives in his shoes, then ASS x-ray everyone's shoes and ban matches.

FAA:
For each dollar an attacker spends, FAA spends a thousand dollars.
There are no easy solutions.
FAA federalizes checkpoints.
FAA brings in more manpower.
FAA brings in more technology.

TSA:
We plan responses in advance.
Nobody needs a metal object to bring down an airliner.
Nobody needs even explosives to bring down an airliner.
Everything on an airplane burns.
Passengers not burn.
Technologies detect oxydizers.
Screeners learntodetect oxydizers.

FAA:
Airlines take the lead on aviation security.
Airlines marketed cheap tickets.
Airlines passed security off on the federal government.
Security officers be on every flight.
Retrain flight attendants as security officers.
We cannot give passengers soda and peanuts.
Passing around soda and peanuts should be secondary.

ASS:
A lot of airlines are not doing well.
A lot of airlines are on government assistance.
Airlines raise prices.
Airlines mishandle baggage.
The TSA changes screening rules constantly.
Anything radical costs a lot of money.
Anything radical deters people.
An economic threat is also a threat.

TSA:

TSA enforce consistency in regulations.
TSA enforce consistency in regulations' application.
Airline bankruptcy has advantages.
Bankruptcy makes it easier to rearrange company assets.
Bankruptcy makes it easier to renegotiate vendor contracts.
Bankruptcy makes it easier to renegotiate supplier contracts.

FAA:
TSA, FAA, and ASS have productive discussions.
TSA, FAA, and ASS get back to work.
TSA, FAA, and ASS come up with concrete measures.
TSA, FAA, and ASS generate some ROI.
The above examples are not all-inclusive.

In the end, after obtaining sentences that are acceptable to Dowser, we decided not to try to analyze the generated OOAM. In any case, the sentences do not even begin to form a correct, complete, and consistent SRS. So there is little point in discovering ambiguities. The set of sentences itself is ambiguous because there are unbounded ways to complete them into a SRS. Indeed, Dowser was never intended to be used in the part of RE in which this sort of raw requirements information is analyzed.

Nevertheless, the exercise of rewriting the sentences to be acceptable to Dowser was valuable in making the intent of the sentences clearer to any reader. It forced us to identify the doers of each activity that was specified in a passive sentence. It forced us to grapple with the meaning of many a phrase that is used in normal conversation that really have no effect on any OOAM. It forced us to find simpler ways to express things that had been expressed in the normal sloppy way of everyday conversation. The result was, in our opinion, a much cleaner set of sentences. Learning how to write these cleaner sentences may be the most valuable effect of the use of Dowser.

This exercise exposed some limitations of the current version of Dowser. Note that there are two sources of limitations, (1) what the link grammar can parse and (2) what the Dowser rules can handle to construct an OOAM from the results of the parse. It was fairly easy to rewrite the sentences so that they could be accepted by the parser. It was more difficult to get Dowser to handle parsed sentences. In each case that Dowser did not accept a construct, the choices were (a) to modify existing Dowser rules or to add new Dowser rules to deal with the construct or (b) to change the containing sentence to avoid the construct. If we could see the effect on OOAMs of the construct, we took the first choice. If not, we took the second choice.

The two main limitations that we discovered are of the second kind. The current version of Dowser cannot deal with modal verbs and with negations, e.g., not. Because we could not see their effect on OOAMs, we would take the second choice response. Because we did not analyze the model, this response was applied only for the modal verbs in the last version of the sentence. That is, the nots are still there. One possible treatment of a not is to build the normal graph with the verb labeling the arc and putting a not by the verb to indicate a misuse case [33].

## 5 Discussion and Conclusion

Dowser, a tool built mostly out of existing software, is able to help a human analyst to identify ambiguity, inconsistency, and underspecification in a NL SRS.

Of course, Dowser's lack of perfection, particularly in the construction of an OOAM and in the DST recall, says that Dowser can be used as only one of an array of approaches and tools for identifying DSTs and for detecting ambiguity, inconsistency, and underspecification in NL SRSs. However, because of the inherent difficulty of these tasks for humans, every little bit helps!

One drawback of the approach is that for best results, the input should be written in the constrained language. If the actual input is not written in the constrained language, it must be rewritten. This rewriting necessity might be considered a reason not to use Dowser. However, one could argue that the rewriting is part of the whole process of eliminating ambiguity, which ultimately the human carries out.

### 5.1 Future Work

The lack of perfection says that more work is needed:

- How does Dowser perform on larger, industrial-strength NL SRSs? Answering this question would help to explore the problem space and to find new unsolved research questions.
- The current Dowser cannot resolve anaphora. An anaphor is a linguistic unit, such as a pronoun, that refers to a previous unit. In The customer can buy text books and return them., them is an example of an anaphor, which must be resolved to text books. While Dowser can identify anaphora, it cannot resolve them. A simple solution would be to have Dowser flag all anaphora in its input text, so a human analyst could change each to its resolution.
- DST identification can be improved. As mentioned above, syntactic information is not sufficient to detect all the DSTs within a document. Therefore, frequency analysis or baseline text analysis [21] might improve DST identification.
- Additional semantic knowledge could improve the capability of Dowser. For example, the *WordNet* lexicon contains information about hypernyms, which can indicate superclasses, and meronyms, which can indicate aggregations. This information could be used to supply missing links in an OOAM. However, although *WordNet* is a large online lexicon, it lacks DSTs and therefore might be only a little help. Extending Dowser's dictionary with DSTs could reduce this problem.
- The current Dowser is not applicable to a NL SRS that has a functional language style, i.e., with sentences such as, The system must provide the functionality of.... Handling such sentences would require a different grammar. Future work could examine which grammar is the most suitable for class extraction.
- The UML offers a set of different diagram types. NLP could be used to create sequence, state, or other diagrams. For example, Juristo *et al.* [17] developed also a controlled grammar for specifying dynamic behavior.
- Other work has found different sources of ambiguities in NL SRS. Since there seems not to be a single perfect approach, different approaches (e.g. [41, 7, 23])

could be integrated into a single framework for validating NL SRSs. This integration could lead to a new method of developing NL SRSs. Indeed, this sort of integration would deal with any syntactic ambiguities found by the parser that is used.

## Acknowledgments

The authors thank the anonymous referees for some really helpful suggestions. Daniel Berry's, work was supported in part by Canadian NSERC Grant Number NSERC-RGPIN227055-00.

## References

[1] Abbott, R. J.: Program Design by Informal English Descriptions. Comm. ACM, 26, 882–894 (1983)

[2] Berry, D. M., Daudjee, K., Dong, J., Fainchtein, I., Nelson, M. A., Nelson, T.: User's Manual as a Requirements Specification: Case Studies. Requir. Eng. J., 9, 67–82 (2004)

[3] Berry, D. M., Kamsties, E.: Ambiguity in Requirements Specification. In: Leite, J. C. S. P., Doorn, J. (eds.) Perspectives on Requirements Engineering, pp. 7–44, Kluwer, Boston, MA, USA (2004)

[4] Bucchiarone, A., Gnesi, S., Pierini, P.: Quality Analysis of NL Requirements: An Industrial Case Study. In: 13th IEEE International Conference on Requirements Engineering (RE'05), pp. 390–394, IEEE Comp. Soc. Press, San Diego, CA, USA (2005)

[5] Clayton, R., Rugaber, S., Wills, L.: Dowsing: A Tools Framework for Domain-Oriented Browsing Software Artifacts. In: Automated Software Engineering Conference, pp. 204–207, Honolulu, HI, USA (1998)

[6] Delisle, S., Barker, D., Biskri, K.: Object-oriented Analysis: Getting Help from Robust Computational Linguistic Tools. In: Friedl, G., Mayr, H. C. (eds.) Application of Natural Language to Information Systems, pp. 167–172, Oesterreichische Computer Gesellschaft, Vienna, Austria (1999)

[7] Fabbrini, F., Fusani, M., Gnesi, S., Lami, G.: The Linguistic Approach to the Natural Language Requirements, Quality: Benefits of the use of an Automatic Tool. In: 26th Annual IEEE Computer Society–NASA GSFC Software Engineering Workshop, pp. 97–105, IEEE Comp. Soc. Press, San Diego, CA, USA (2001)

[8] Fuchs, N. E., Schwertel, U., Schwitter, R.: Attempto Controlled English (ACE) Language Manual, Version 3.0. Tech. Rept. 99.03, Dept. Computer Science, U. Zurich (1999)

[9] Fuchs, N. E., Schwitter, R.: Attempto Controlled English (ACE). In: 1st International Workshop On Controlled Language Applications (CLAW), pp. 124–136, Leuven, Belgium (1996)

[10] Gause, D., Weinberg, G.: Exploring Requirements: Quality Before Design. Dorset House, New York, NY, USA (1989)

[11] Gemini Natural-Language Understanding System, http://www.ai.sri.com/natural-language/projects/arpa-sls/nat-lang.html
[12] Gervasi, V., Nuseibeh, B.: Lightweight Validation of Natural Language Requirements. Softw., Pract. & Exper., 32, 113–133 (2002)
[13] Graphviz – Graph Visualization Software Home Page, http://www.graphviz.org/Credits.php
[14] Grishman, R.: Information Extraction: Techniques and Challenges. In: Pazienza, M. T. (ed.) Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, SCIE'97: International Summer School on Information Extraction, LNAI, vol. 1299, pp. 10–27, Springer, London, UK (1997)
[15] Harmain, H. M., Gaizauskas, R. J.: CM-Builder: A Natural Language-Based CASE Tool for Object-Oriented Analysis. Autom. Softw. Eng., 10, 157–181 (2003)
[16] Heimdahl, M.: An Example: The Lift (Elevator) Problem. http://www-users.cs.umn.edu/heimdahl/formalmodels/elevator.htm
[17] Juristo, N., Moreno, A. M., Lopez, M.: How to Use Linguistic Instruments For Object-Oriented Analysis. IEEE Softw., 17, 80–89 (2000)
[18] Kiyavitskaya, N., Zeni, N., Mich, L., Berry, D. M.: Requirements for Tools for Ambiguity Identification and Measurement in Natural Language Requirements Specifications. Requir. Eng. J., 13, 207–239 (2008)
[19] Klein, D., Manning, C. D.: Accurate Unlexicalized Parsing. In: 41st Meeting of the Association for Computational Linguistics, pp. 423–430 Assoc. for Computational Linguistics, Morristown, NJ, USA (2003)
[20] Kof, L.: Natural Language Processing for Requirements Engineering: Applicability to Large Requirements Documents. In: Russo, A., Garcez, A., Menzies, T. (eds.) Workshop on Automated Software Engineering, Linz, Austria (2004)
[21] Lecoeuche, R.: Finding Comparatively Important Concepts between Texts. In: 15th IEEE international Conference on Automated Software Engineering, pp. 55–60, IEEE Comp. Soc. Press, San Diego, CA, USA (2000)
[22] Luqi, Kordon, F.: Case Study: Air Traveling Requirements Updated (Blog scenario). In: this volume Springer, Berlin, Germany (2008)
[23] Mich, L.: On the Use of Ambiguity Measures in Requirements Analysis. In: Moreno, A., van de Riet, R. (eds.) 6th International Conference on Applications of Natural Language to Information Systems (NLDB'01), pp. 143–152, LNI, vol. 3, Gesellschaft für Informatik, Bonn, Germany (2001)
[24] Mich, L., Franch, M., Novi Inverardi, L.: Market Research for Requirements Analysis Using Linguistic Tools. Requir. Eng. J., 9, 151–151 (2004)
[25] Mich, L., Garigliano, R.: Ambiguity Measures in Requirements Engineering. In: International Conference on Software–Theory and Practice (ICS2000), 16th IFIP World Computer Congress, pp. 39–48, Publishing House of Electronics Industry, Beijing, China (2000)
[26] Mich, L., Garigliano, R.: NL-OOPS: A Requirements Analysis Tool Based on Natural Language Processing. In: 3rd International Conference on Data Mining Methods and Databases for Engineering, Witpress, Southampton, UK (2002)
[27] Miller, G. A., Felbaum, C., *et al.*: WordNet Web Site. Princeton U., Princeton, NJ, USA, http://wordnet.princeton.edu/

[28] Mollá, D., Schwitter, R., Rinaldi, F., Dowdall, J., Hess, M.: ExtrAns: Extracting Answers from Technical Texts. IEEE Intelligent Syst. 18, 12–17 (2003)
[29] Nanduri, S., Rugaber, S.: Requirements Validation via Automated Natural Language Parsing. J. Mgmt. Inf. Syst. 12, 9–19 (1995–96)
[30] Rumbaugh, J.: Object-Oriented Modeling and Design. Prentice Hall, Englewood-Cliffs, NJ, USA (1991)
[31] Rupp, C., Goetz, R.: Linguistic Methods of Requirements-Engineering (NLP). In: European Software Process Improvement Conference (EuroSPI), Copenhagen, Denmark, http://www.iscn.com/publications/#eurospi2000 (2000)