Validation of STEP/EXPRESS Specifications by Automatic Natural Language Generation

Hercules Dalianis, Anders Hedman and **Paul Johannesson**

Department of Computer and Systems Sciences
Royal Institute of Technology and Stockholm University
Electrum 230, 164 40 Kista, Sweden
{hercules, ahedman, pajo}@dsv.su.se

Abstract

The STEP/EXPRESS standard is one of the most complex ever developed. Understanding its workings is not an easy task for new users. Designers and other persons involved in the modelling process need to validate STEP/ EXPRESS specifications. They need, therefore, a tool generating natural language descriptions from STEP/EXPRESS specifications.

This paper investigates the differences between the STEP domain and other domains with respect to natural language generation architectures. In order to do this, a questionnaire was handed out to personnel at Volvo Data and DSV-KTH-SU asking how to paraphrase parts of a PDM (Product Data Model) into Natural Language text.

The texts were then analysed using aggregation and discourse techniques. The findings were implemented in a tool that automatically generates natural language descriptions from STEP/EXPRESS specifications derived from Application Protocol 214 *Core Data for Automotive Design Process*.

1. Introduction

1.1. The promise of new ways of working

One important technology to combat the isolation of computer systems is ISO 10303 (STEP) (Al-Timimi & MacKrell 96), which is an international standard for the representation and exchange of product model data. for the manufacturing industry.

Data models of STEP are formally specified in the language EXPRESS (ISO-91, Schenk & Wilson 94). EXPRESS is a static modelling language of entity-relationship type. EXPRESS provides also instance models which are called Step Physical Files (SPF).

1.2 . Understanding and using STEP

Product development using STEP can be both tedious and cumbersome, not only is the actual writing of code time consuming, but checking its validity can also take a considerable amount of time and effort. Furthermore, there are large groups of people who have little knowledge of STEP and can not accurately interpretate the data models. Thus, we see a potentially very large group of people who need more efficient tools to aid them in data model interpretation as well as validation.

1.3 . The aim of this paper

Our contribution is to empirically investigate the requirements for building a natural language generation (NLG) system for STEP models. These requirements will be a basis for developing a natural language generation system for paraphrasing specifications expressed in the STEP/ EXPRESS standard, in particular from specifications using the STEP Application Protocol 214 Core Data for Automotive Design Process.

The requirements for validating by means of NLG are of course different for STEP models compared to other domains as e.g. automatic weather reports, car manuals, medical informatics, expert systems explanations and machine translation.

The approach was to investigate schemas as well as instances, expressed in STEP/EXPRESS and by using a questionnaire on Automotive Designers and Constructors at Volvo as well as computer scientists at DSV-KTH-SU (Department of Computer and Systems Science, Royal Institute of Technology and Stockholm University) to obtain the "correct" natural language expressions describing the formal specifications. These specifications were analyzed according to discourse structure, semantic content, syntactic structure and lexical choice.

Based on this we implemented a Sentence Planner and a Natural Language Surface Grammar and Lexicon in Prolog for the generation of Natural Language (English).

The contribution of this project is that we will create a support tool for the STEP/EXPRESS

This work was funded by Volvo Research Foundation

standard which will help designers and constructors of cars to understand and correct their EXPRESS specifications by reading Natural Language output. This support tool will not only help designers and constructors but also other persons involved in the car design process to validate the formal specification by reading it in natural language.

2. Discussion

2.1. Validation and explanation

The problem of validating models, i.e. ensuring that they correspond to the stakeholders' intentions, is a well-known problem within the software community. Errors and misunderstandings identified early in the development process are in general simple to remove. Validating STEP schemata is problematic mainly for the following reasons:

STEP schemas are conceptual models and are not executable. This means that the correctness of a STEP schema, as part of an implementation, can be validated only by means of theoretical analysis.

Within the process of developing STEP schemata, there are two different kinds of STEP users. The first kind are the STEP end users who are specifying requirements for STEP schemata. The second kind is the STEP developers who are incorporating these requirements into the design of STEP schemata. The STEP end users have problems with checking whether and how their requirements have been incorporated, i.e. they have problems in validating the STEP schemata. There are several reasons for this. First, end users are often not sufficiently familiar with the EXPRESS language. Secondly, STEP schemata have little formal semantics so that a STEP user has too much freedom in the interpretation of STEP schemata. Thirdly, STEP schemata are very complex and extensive. Therefore, it is difficult to obtain an overview of the content of STEP schemata without supporting techniques and tools.

An other method to validate EXPRESS schemas is to represent them in EXPRESS-G for better readability.

2.3. EXPRESS-G

EXPRESS-G provides a graphical representation of EXPRESS. Although such graphical representations can be of considerable use for someone working with STEP, they are not always sufficient for validation purposes. One apparent problem is that while EXPRESS-G models are considerably more accessible than EXPRESS, they still require plenty of background understanding on the interpreter's part. Moreover, EXPRESS-G only covers a subset of EXPRESS, thus there are sentences of EXPRESS which can not be visualized in EXPRESS-G.

Our approach is to accept both the benefits and the limitations of EXPRESS-G. Our NLG-tool should be viewed as a complement to EXPRESS-G.

3. Related research

Some work has been carried out previously in the area of validation of Entity-Relationship models by natural language generation (NLG), but never specifically in the area of validation of STEP/ EXPRESS models by NLG.

Arguments to use natural language generation for validation of formal specifications are presented in (Swartout 82). A set of translation rules for translating Entity-Relationship diagrams to natural languages (NL) was defined in (Chen 83). One of the first attempts to generate natural language from a conceptual model was the AMADEUS system described in (Black 87).

One ambitious project to create a support tool to specificy telecom services through the whole requirement engineering process was VINST (VIsual and Natural language Specification Tool). The VINST tool was constructed at the telecom company Ericsson. In the VINST tool the user could specify telephone services with pictures and natural language and obtain feedback of the specification by natural language generation and execution of the ready specification. The VINST tool is described in (Bretan et al 95) and VINST's Natural Language generator is described in (Dalianis 95). The generation of non-redundant (aggregated) natural language descriptions of formal specifications are described in (Dalianis & Hovy 96, Dalianis 96).

Aggregation is the process of removing redundant information without changing the information content in a text. In linguistics analysis aggregation is called ellipsis and coordination.

A suggestion to generate a whole Natural Language discourse built on Hobbs coherence relations (Hobbs 85) for validation of a conceptual model was made in, (Dalianis 92) in augmentation to single sentence generation.

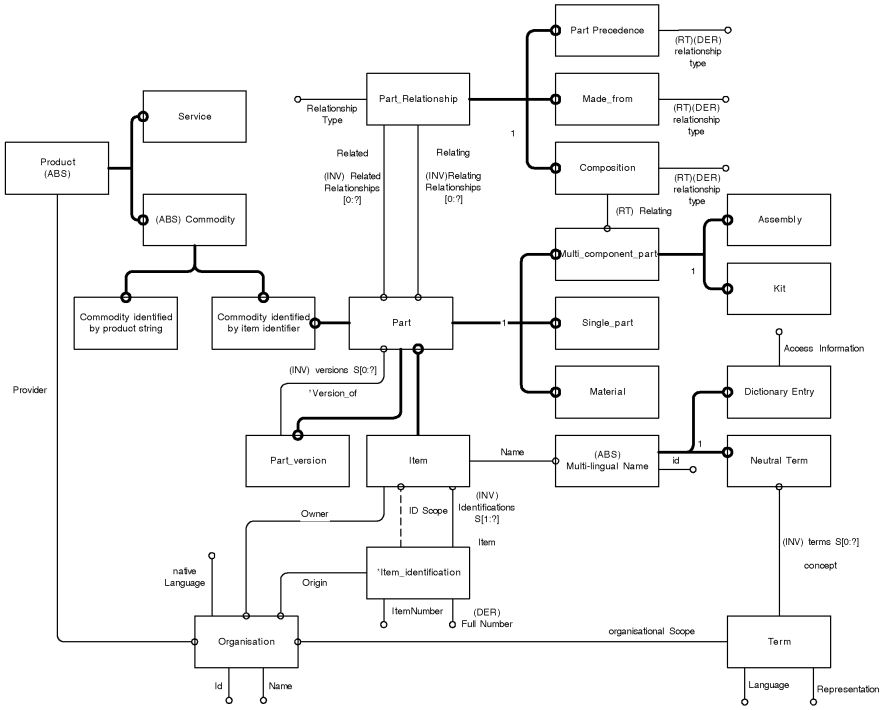

Figure 1. PDM model in EXPRESS-G expressing "Part" from Volvo

A discourse is a set of coherent sentences.

A system design for an explanation component for conceptual modelling based on (RST Rhetorical Structure Theory), (Mann & Thompson 88), is described in (Gulla 96).

4. Questionnaire, Analysis and Implementation

4.1. Questionnaire

To identify texts appropriate for validation of STEP models we had to gather a corpus. To carry out this, we constructed a questionnaire to collect texts describing STEP models. The questionnaire contained a PDM (Product Data Management) requirements model expressed in both textual EXPRESS and EXPRESS-G, parts of it were enlarged and questions were asked relating various entities and their relationships. The answers should be in Natural Language English. The questionnaires contained also syntax descriptions of EXPRESS-G and a dictionary of the terms used in the PDM-model (see previous section Figure 1).

The persons who filled in the questionnaires answered questions pertaining to the EXPRESS-G diagrams. More specifically, they were asked to identify a selected number of entities and write descriptions of them. In producing the descriptions they were allowed to use their own terms as well as those in the dictionary.

The PDM-model was obtained from Volvo Data and contained a model describing the company's PDM-requirements, specifically the parts descriptions within five of the different Volvo companies, together with a dictionary with definition of terms.

Sixteen questionnaires were sent out to personnel at Volvo Data (Ten questionnaires) and DSV-KTH-SU (Six questionnaires) of these nine were filled in and returned to us.

4.2. Analysis

The analysis of the answers in the questionnaires were based on the following criteria: Were the answers of Discourse or Single sentence structure? How many sentences were used? Was dictionary information used in the discourses? (i.e. information from the dictionary not available in the PDM-model) Which aggregations were used? see (Dalianis & Hovy 96). How were cue words used (i.e. disambiguation of aggregation ? Was referring expressions (pronouns etc.) used? and finally;

The reason to use these criteria is that they have been proven to be valuable when analysing texts which will later will be used as a basis for automatic natural language generation. (Dalianis & Hovy 96, Dalianis 96)

Table 1. Describes all the findings from the questionnaires

Person no	Discourse	Dictionary information	Sentences	Aggregation	Cue words	Referring expressions	Missing answers
1	5	5	22	4	2	0	10
2	4	3	13	1	1	0	11
3	3	4	22	2	0	1	8
4	7	7	36	5	1	3	4
5	6	8	28	3	0	1	6
6	0	5	16	0	0	0	14
7	1	8	14	3	0	2	12
8*)	13	3	27	8	0	1	9
9	4	7	22	3	2	1	7
Sum:	43	50	200	29	6	9	81

Questions asked	252
Missing answers †)	81
Obtained answers	171

†) We guess that people got too tired to answer all the 28 complicated questions in the questionnaire

*) Uses a lot of domain and also instances information, see below.

An example below on a representative natural language discourse (from the questionnaire).

How is an assembly related to a kit?

(1) Assemblies and Kits are both entities
(2) They are subtypes of a Multi_ Component_Part
(3) A Kit is put together for temporary use
(4) An Assembly is put together such that it is ready for its intended main use

What we observe is a natural language discourse containing four sentences. In sentence (1), we see that Predicate-Subject aggregation and the cue word *both* were used. Sentence (2) is a referring expression through the use of the pronoun *they*. Sentences (3) and (4) use dictionary information which was not available neither in the EXPRESS-G model nor in textual EXPRESS.

We carried out an analysis of the texts from the questionnaires and found, as expected, that the texts were aggregated and discourses were used in many cases. A surprising finding, though, was that people relied heavily on dictionary information to describe the model. This dictionary information was not contained in the model. Furthermore, people experience a need to use instances to exemplify/illustrate their different answers.

From the analysis we found also that the ISA-relation were used as the first discourse element and the dictionary information came at the end of the discourse.

One of the answers, the one from a real domain expert, (See Table 1, above, Person no 8, The domain expert), used domain information heavily which neither was available in the EXPRESS model nor in the dictionary. See sentences (3) and (4) above and also sentences and enumeration of words like (*Kit is used for repair, sale, transport. Material: Bar, hose, pipe, fluid, powder. Parts Fixed: welded, glued, reveted. Detachable: screwed, damped etc*).

Since EXPRESS is a static modelling language, one has to model the temporal constructs in a static way by using some sort of entity type description. In the example sentences (3) and (4), this has not been done in the STEP model, so to distinguish Kit and Assembly the users in their answers utilized dictionary information.

Our general feeling is that the domains modelled in STEP are not very constrained or wellknown by their modellers. Therefore the STEP models are very superficial in respect to their domain.

Moreover STEP Physical Files (SPF) (instances) are very difficult to read and understand. SPF is not tested in this investigation since the obtained SPF from Volvo Data was not complete.

Our recommendation is therefore to use canned text extracted from the dictionary in the NLG system, and eventually also use instances as illustrative examples.

4.3. Implementation

ASTROGEN (**A**ggregated deep and **S**urface na**T**u**R**al language **GEN**erator) is written in Prolog and was developed during various research project, see (Dalianis 96), specifically in collaboration with Ericsson telecom company. ASTROGEN takes as input a part of a formal specification and translates it to a natural language English text. This is carried out with the help of a sentence planner, a surface grammar and a lexicon (see Figure 2, below). '

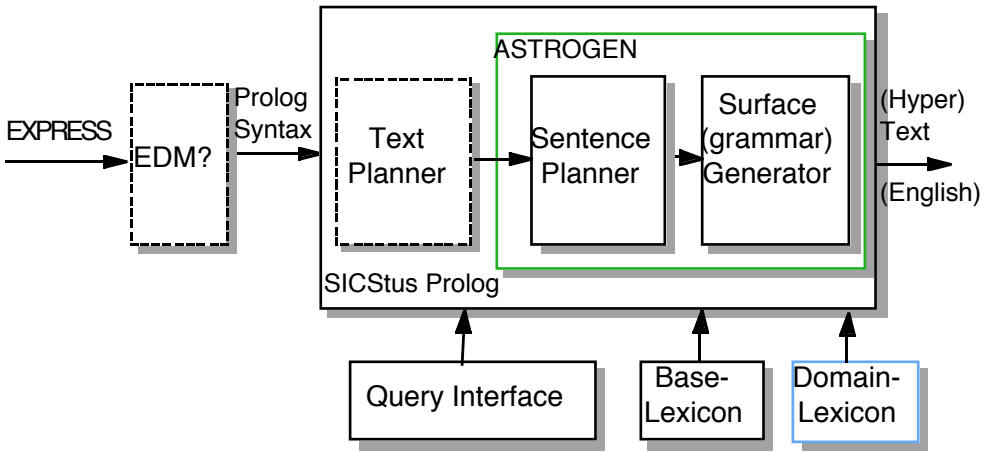

Figure 2. Architecture of the ASTROGEN Natural Language Generator

ASTROGEN has been adapted to the new domain by constructing a new domain lexicon for the PDM terms. (39 new lexical items). The sentence planner used only two aggregation rules the Subject-Predicate and the Predicate-Direct Object aggregation rules. ASTROGEN has also been enhanced with the capability to include dictionary information in the form of canned text at the end of the discourse .

From the questionnaires we found that dictionary information was used and therefore we needed to include the feature to incorporate canned text into the generated discourses. Furthermore we extended the ASTROGEN surface generator with the capability to delimit sentence clauses with commas instead of "and". The example below shows the first draft implementation where we have handcoded part of the Volvo PDM model into Prolog.

(Normal generation with no features)
?- question(kit&assembly).
a Multi_Component_Part is a supertype for a Kit and
a Kit is an entity and
a Kit is a subtype of a Multi_Component_Part and
a Multi_Component_Part is a supertype for an Assembly and
an Assembly is an entity and
an Assembly is a subtype of a Multi_Component_Part.
yes
(All aggregation rules)
?- all_rules.
yes
?- question(kit&assembly).
an Assembly and a Kit are entities and
an Assembly and a Kit are subtypes of a
Multi_Component_Part and
a Multi_Component_Part is a supertype for an Assembly and
a Kit.
yes
(Pronominalization)
?- pronoun.
yes
?- question(kit&assembly).
an Assembly and a Kit are entities and
they are subtypes of a Multi_Component_Part and
a Multi_Component_Part is a supertype for an Assembly and
a Kit.
yes
(Clause delimitation by a comma)
?- clause_comma.
yes
?- question(kit&assembly).
an Assembly and a Kit are entities ,
they are subtypes of a Multi_Component_Part ,
a Multi_Component_Part is a supertype for an Assembly and
a Kit.
yes
(Use dictionary information as canned text)
?- dictionary_pred.
yes
?- question(kit&assembly).
an Assembly and a Kit are entities ,
they are subtypes of a Multi_Component_Part ,
a Multi_Component_Part is a supertype for an Assembly and
a Kit.
A Kit is put together for temporary use
An Assembly is put together such that it is ready for its
intended main use
yes

The improvements we performed on ASTROGEN to adhere to the findings in our questionnaires were: To add a module for interleaved generation of canned text (that means hard-coded text) with the text originating from the STEP model. Furthermore, a module was implemented which did comma-separation of the different clauses except from the aggregated ones.

5. Conclusions and future directions

Our main finding is that the STEP model investigated was superficial in respect to the domain modeled. Therefore, the test persons used heavily dictionary information to explain the domain, and specifically the domain expert used also instances that were not available in the questionnaire. Therefore we recommend the use of canned text in the automatically generated discourses.

One more general proposal for the STEP domain to support modelling is by using an ontology to obtain a better domain knowledge and support as carried out in the electrical domain in (Dalianis & Persson 97). We will investigate other domains, Application Protocols (APs), e.g. for ships, electrotechnical plants etc, and propose guidelines for how to create lexicons for other domains with minimal work by reusing the results from this work.

One strength of this paper is the synergy effect: that other domains (other APs) in the STEP/ EXPRESS world could make use of our results and our guidelines to easily create natural language generation systems.

The parts in ASTROGEN (see Figure 2.) which are still missing is the translator from EXPRESS textual format to the Prolog format ASTROGEN requires, and a text planner to make it possible to create discourses. For the translator from EXPRESS to Prolog we are planning to use a commercial tool called EDM (Express Data Manager) from EPM Technologies, and for the discourse planner we intend , see (Dalianis & Johannesson 97) to implement a text planner based on Toulmin's argumentation model (Toulmin et al. 79) enhanced with RST primitives (Mann & Thompson 88).

Acknowledgements

We would like to thank for the kind assistance of people at Volvo Data and also the computer scientists at DSV-KTH-SU for taking the time to answer questions and meet with us. We would especially like to thank Per Brorson and Stefan Lindahl at Volvo Data who have proved to be great discussion partners willing to share their expertise.

References

(Al-Timimi & MacKrell 96) K. Al-Timimi and J. MacKrell: STEP: Towards Open Systems. STEP Fundamentals & Business Benefits, CIMdata, 1996.

(Black 87) W.J.Black. Acquisition of Conceptual Data Models from Natural Language Descriptions, In The Proceedings of The Third Conference of the European Chapter of Computational Linguistics, Copenhagen, Denmark 1987.

(Bretan et al. 95) I. Bretan, M. Engstedt & B. Gambäck: A Multimodal Environment for Telecommunication. Specifications.In Proceedings of the 1st International Conference on Recent Advances in Natural Language Processing, pp. 191-198, Tzigov Chark, Bulgaria, September, 1995.

(Chen 83) P. P-S. Chen: English Sentence Structure and Entity Relationship Diagrams, Information Sciences 29, p.p. 127-149, 1983.

(Dalianis 92) H. Dalianis: A method for validating a conceptual model by natural language discourse generation. CAISE-92 Int. Conf. on Advanced Information Systems Engineering, Loucopoulos P. (Ed.), Springer Verlag Lecture Notes in Computer Science, no 593, pp. 425-444, 1992.

(Dalianis 95) H. Dalianis: Aggregation in the NL-generator of the VIsual and Natural language Specification Tool. In Proceedings of The Seventh International Conference of the European Chapter of the Association for Computational Linguistics (EACL-95), Student Session, pp 286-290, Dublin, Ireland, March 27-31, 1995.

(Dalianis 96) H. Dalianis: Concise Natural Language Generation from Formal Specifications, Ph.D. dissertation, (Teknologie Doktorsavhandling), Department of Computer and Systems Sciences, Royal Institute of Technology/Stockholm University, June 1996, Report Series No. 96-008, ISSN 1101-8526, SRN SU-KTH/DSV/R--96/8--SE.

(Dalianis & Hovy 96) H. Dalianis, and E.H. Hovy. Aggregation in Natural Language Generation.,*Trends in Natural Language Generation: an Artificial Intelligence Perspective*, EWNLG'93, Fourth European Workshop, G. Adorni & M. Zock (Eds), Springer Verlag Lecture Notes in Artificial Intelligence no 1036, pp 88-105, 1996

(Dalianis & Persson 97) H. Dalianis & F. Persson: Reuse of an ontology in an electrical distribution network domain. Presented at the AAAI 1997 Spring Symposium Series, Ontological Engineering March 24 - 26, 1997, Stanford University, California.

(Dalianis & Johannesson 97) H. Dalianis & P. Johannes son: Explaining Conceptual Models -- An Architecture and Design Principles, to be presented at the 16th International Conference on Conceptual Modeling (ER'97),Los Angeles, California, USA, Nov 3 - 6, 1997

(Gulla 96) J. A. Gulla: A General Explanation Component for Conceptual Modeling in CASE Environments, ACM Transcation of Information Systems Vol 14. No 3. pp 297-329, July 1996.

(Hobbs 85) J.R Hobbs: On the Coherence and Structure of Discourse, Center for the Study of Language and Information, Report No.CSLI-85-37, October 1985.

(ISO-91) The EXPRESS Language Reference Manual, ISO TC184/SC4/WG5, N14, Leeds, April 29,1991.

(Mann & Thompson 88) W.C. Mann S.A.Thompson Rhetorical Structure Theory: Towards a Functional Theory of Text Organization, In TEXT Vol 8:3,1988.

(Schenk and Wilson 94) D.Schenk and P. Wilson Information Modeling the Express Way, Oxford University Press, 1994.

(Swartout 82) B.Swartout: GIST English Generator: In Proceedings of AAAI-92, American Association of Artifical Intelligence, Carnegie-Mellon University and University of Pittsburgh, Pittsburgh, Pennylvania, 1982.

(Toulmin et al 79) S. Toulmin, R. Rieke and A. Janik: An introduction to reasoning, McMillan Publishing Company, 1979