MULTIDOC – Controlling Language in multilingual Documentation

Jörg Schütz

1. Introduction

MULTIDOC is a European project of the Fourth Framework Programme within the Language Engineering Sector. It is founded on the specific needs and requirements of product documentation expressed by several representatives of the European automotive industry, among them are Bertone, BMW, Jaguar, Renault, Rolls-Royce Motor Cars, Rover, Volvo and others. The focus of the project is particularly on the multilingual aspects of product documentation. Therefore, the general goal is to define and specify methods, tools and workflows supporting stronger demands on quality, consistency and clarity in the technical information, and shorter lead times and reduced costs in the whole production value cycle of documentation including the translation into multiple languages.

The results of the project are applicable to any other component or system manufacturing business; thus, they are not restricted to the automotive industry. The project is divided into two phases: an inception and elaboration phase, the so-called MULTIDOC Concerted Action (LE3-4230), and a construction or development phase, the so-called MULTIDOC Project (LE4-8323). The first phase has been finished by the end of last year, and the second phase has started in January 1998.

2. Basic Requirements and Vision

The aim of the MULTIDOC Concerted Action was to identify the problem areas and to specify solutions for the European automotive industry when it comes to multilingual product documentation and also set a roadmap for the future. In software engineering, this phase is usually called the inception phase of an iterative software development process. During inception we establish the business rationale for the project and decide on the scope of the project. This is also the phase where we get the commitment from the project sponsor(s) to go further; in our case this was the successful evaluation of our MULTIDOC Project proposal.

The Concerted Action also included parts of the elaboration phase of a software development project. In elaboration, we collect more detailed requirements, do high-level analysis and design to establish a baseline architecture, and create the plan for construction which is the actual software production phase consisting of many iterations. In our domain, the most crucial bottlenecks comprise the following business areas that needed further elaboration:

- More and more languages in which product documentation has to be published; there is a drastically increased focus on Asian and East-European markets.
- Increasing costs for translations.
- Lead times in the document production process and in the translation process.
- Poor or no possibility to measure and control the translation process.
- Inconsistent use of information structure and information content.

All project partners agree that besides the quality of the product the services associated with the product and the accompanying documentation of the product must be seen as an integral part of the product. To satisfy the demand for high-quality technical

documentation, the documentation has not only to be comprehensible and up to date, it has to be produced and delivered (including the accessibility to new or up-dated information) with modern technologies. The following scenario shall exemplify the intended direction:

Mr M is the proud owner of a new, environmentally clean car which was assembled according to his wish list from a huge variety of car components of the automotive manufacturer. The first contact with his new car was well before the actual delivery in a virtual reality animation, where Mr M was able to check his colour selection, the harmonisation of the chosen colour with the selected materials of the car interior, as well as first virtual driving tests.

Mr M is also very satisfied with the delivered car documentation; he got the personalised documentation right after the deal was contracted. This documentation is not only personalised but also customised to his specific car: this includes the appropriate colouring of all graphics in the documentation and the text itself, where we do not find any generalised references such as "... applies to specific countries.", "... according to the model variant." and so forth.

With the CD-ROM edition of the documentation Mr M can directly search for information on his PC at home, and he is also provided with multimedia enhanced information about his car. This documentation is also available in his car via the on-board computer. This computer maintains each service work and possible repair work in its storage, and therefore permits customised service and repair measures and fault tracing at dealer's workshops. Not to mention, that these data are also available via the world-wide computer network of the car manufacturer.

Mr M gets new information about his car and about new products and services of the car manufacturer via his Internet access at home; for this he has subscribed to the free information service which in addition is parametrisable according to his information demands.

Now, on a business trip to the south of Spain Mr M has to stop with a defect late at night. The workshop that is alarmed due to the 24-hours assistance service is somehow lost: this kind of defect is not listed in the technical service documentation and therefore not a standard service and repair routine. However, the immediate access to the hotline information service via the computer network identifies the same defect two days ago in Oregon/USA. The problem solution that the Spanish mechanic gets on his workshop screen is in English. So he activates the translation-on-demand button and receives a Spanish translation within a few seconds. This translation is not perfect, but the necessary repair steps and the terminology of the needed tools and parts is correct due to a multilingual terminology knowledge base which is maintained by the car manufacturer; therefore the mechanic does not care about the 'ser/estar' errors of the delivered Spanish text.

Not every service information is available in all languages that are supported by the car manufacturer; the translation of service information is a matter of information need, but every available information is accessible through "pull" technology (for example, the hotline information service). Information of common interest is distributed in all languages through "push" technology; this ensures a fast and efficient update of all product documentation. The defect of Mr M's car is now available in English and Spanish (after a correction of the computer translation); after an in-depth analysis of the defect this information might become available in all supported languages.

This scenario, which of course is partly a vision, shows the necessity of integrating services, documentation and networked information technology (IT) solutions with the support of modern, multilingual language technology (LT). The identified documentation bottlenecks and parts of this scenario then form the basis of the MULTIDOC vision of an Abstract Documentation Factory (ADF).

3 MULTIDOC Virtual Application

3.1 Translation Engineering

Within the Concerted Action a so-called virtual application was defined. It constitutes a compromise between the present situation of product documentation in the different automotive companies and the MULTIDOC vision of an ADF that is based on the concept of Translation Engineering. We talk about a virtual application because this

application is based on the generalisation of the different processes and workflows maintained in the automotive product documentation environments of the MULTIDOC partners. For this we analysed the existing documentation processing chain and we identified the stages for initial improvements taking into account the historical evolution of product documentation. The sequential stages of today's documentation value cycle is profiled in Figure 1 below.

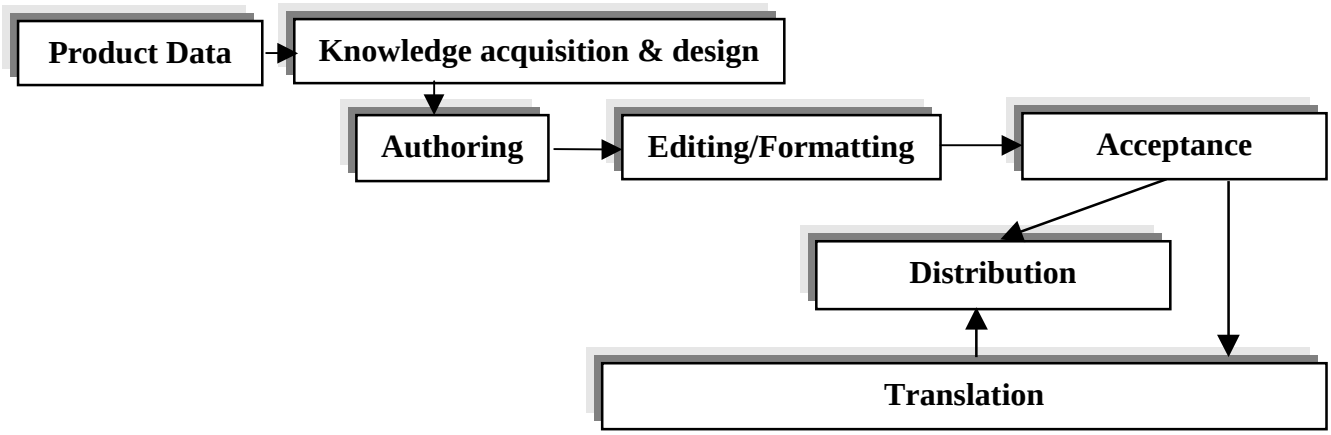

Figure 1: Today's Documentation Workflow

This analysis led to the definition of a strategy for an efficient and effective employment of language technology to bridge the gap between the present situation in product documentation and the MULTIDOC vision. This strategy is based on the present situation and has to be maintained with various restrictions for the different automotive companies but with the common interest to work toward the ADF vision that is shared by all companies, however, with different ways to reach the vision.

The virtual application is the result of the elaboration phase, and it allows for a smooth and cost effective transition of the business, because we have first and foremost concentrated on the existing process stages, where several control capabilities for the source language, such as spell, grammar and style checking functionality in the authoring stage, as well as the control of terminology consistency in the product data stage, the knowledge acquisition stage and the authoring stage, support the technical writer and other knowledge workers in identifying and defining information objects in an SGML authoring environment.

In broad terms, an information object is either a meaningful, non decomposable SGML marked-up text unit or a composition of such text units. The virtual application shall already include steps toward Translation Engineering that is the operational foundation of the MULTIDOC vision. Translation Engineering (TE) as a business strategy is concerned with:

- Fostering the use of information objects preferably linked with product data to ensure the consistent use of information structure and information content.
- Optimising the translation production chain through the employment of different multilingual language technologies, including multilingual generation.
- Linking of source language information and target language information to facilitate better maintenance, quality assurance and quality control.

TE as a methodology for multilingual documentation will help to drastically reduce cost,

to shorten lead times and to further improve the quality of technical documentation. In addition, the core business and the component or system manufacturing industry benefits in terms of:

- accelerating the building of enterprise-wide and industry sector wide knowledge systems (repositories and knowledge bases) based on Web technology (intranets and extranets) including multimedia (text, graphics, video and animation, virtual reality) and multimodal (language and speech navigation) capabilities,
- improving the semantic content of information objects (accuracy and quality),
- speeding up of the translation processes (today, in limited cases the translation process could be substituted by multilingual generation and symbolic authoring, which in particular has to be seen in combination with controlled languages),
- reorganising of the overall production process (lean multilingual documentation).

A snapshot of this information distribution scenario is shown in Figure 2 below.

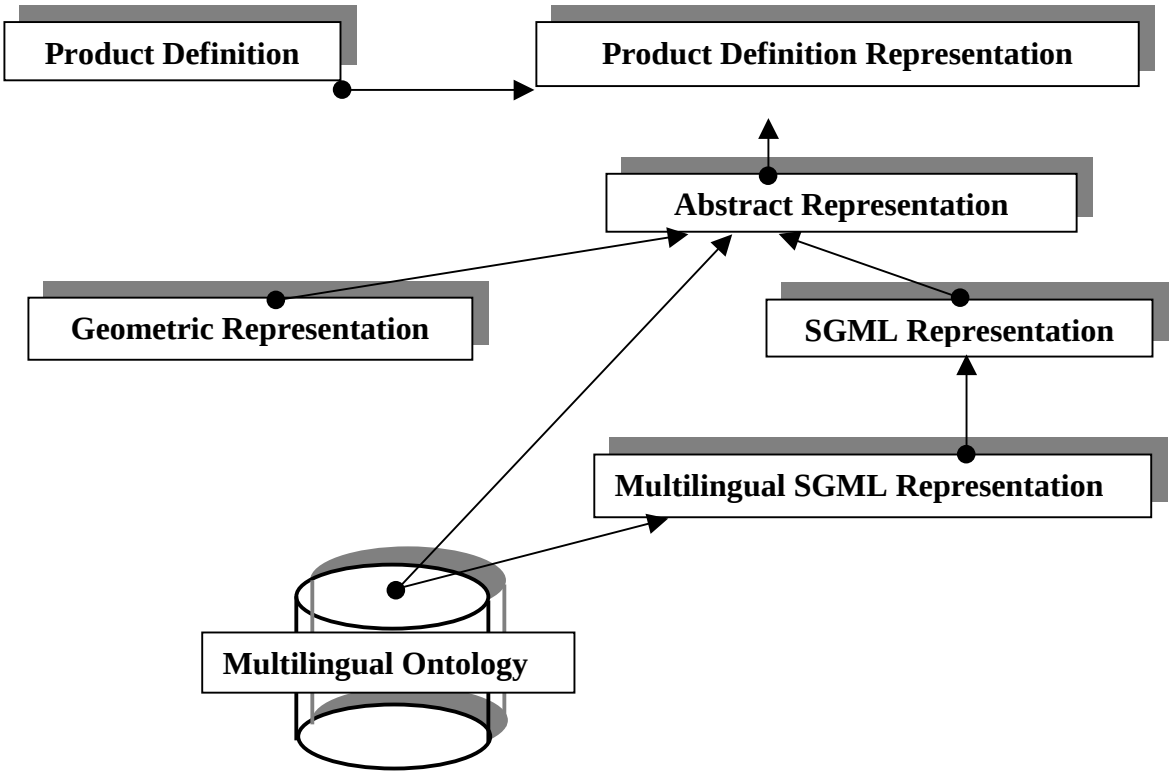

Figure 2: Information Distribution

TE as such will revolutionise the current way of thinking in technical documentation because the whole documentation process is oriented toward multilingualism. This new business scenario includes a push/pull policy for technical information delivery and retrieval in an automotive dealer's workshop in combination with a translation-on-demand policy (see the scenario above). TE is responsive to the new business demands, and it will harmonise and unify the most crucial documentation requirements in areas such as the consistent use of technical information in structure and content, the efficient and effective reuse of information objects based on standardised information structures (increased retrieval hit-rates), and the terminological and multilingual orientation of the whole information production process.

3.2 Abstract Documentation Factory

The vision of the ADF comprises the complete re-organisation of the documentation processes. This means that LT in general, and specifically multilingual LT including translation technology, will make the move from a supporting technology to an enabling technology. The most important areas for this development are:

- Graphics and other multimedia incarnations, such as video, animation and virtual reality applications, may enrich or even replace text in certain information objects and facilitate new approaches to information production such as symbolic authoring.
- Translation-on-demand policy to allow for an efficient and effective control of the actual translation needs, because not all information objects need to be stored in every language that is supported by the business (translation management).
- Compilation of documents from multilingual information objects, either already stored in a foreign language, translated on demand, or generated from an abstract representation; this allows for the simultaneous delivery (publishing) of multilingual documentation.

The focus of the ADF is on the following three main components:

- Multilingual terminological ontology as a means for representing domain knowledge (the subject of technical documentation) linked with natural language semantics. Figure 3 below shows the different views on the ontology organisation at the upper level.
- Object Modelling Technique (OMT) as a theoretical foundation for analysis and design, and as an implementation platform based on distributed object environments such as, for example, CORBA.
- Agent technology as the overall umbrella for construction, and as an alternative implementation platform, especially for networked applications.

In the ADF, knowledge producer and knowledge consumer will operate in virtual environments brokered by software agents. A software agent acts autonomously on behalf of a person to fulfil the person's goal or task. Agents are also key enabler for push technology which is used in information update tasks and information retrieval tasks.

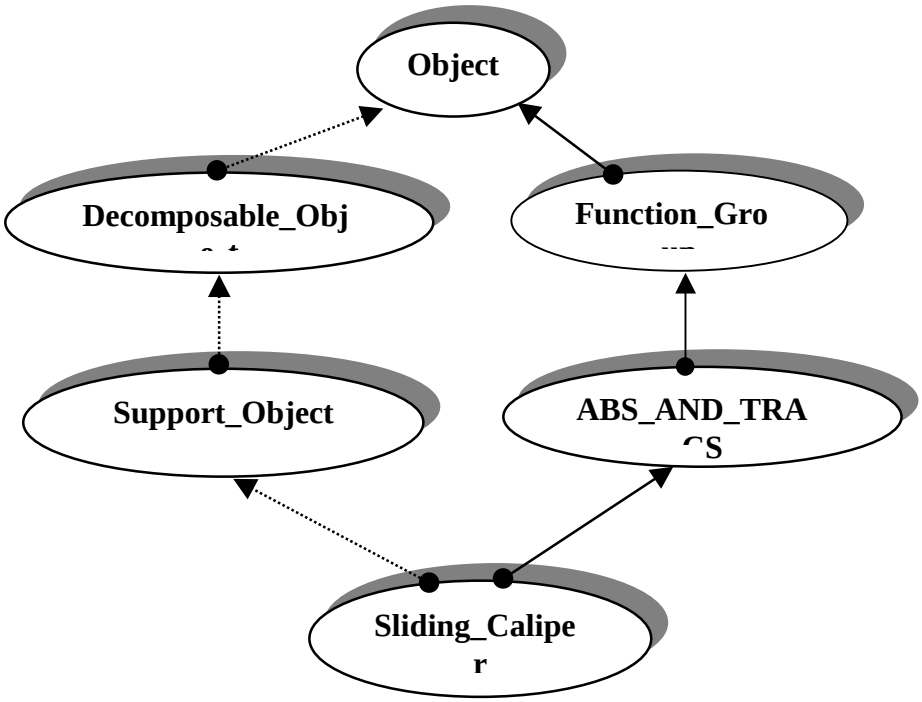

Figure 3: Different Views on the Ontology

3.3 Validation

All development strategies have been validated with a cost/benefit appraisal based on a hypothetical business calculation of a virtual automotive enterprise. We have taken this way to further maintain the generalisation direction which we already followed in the other phases of the Concerted Action. However, our profitability assessment is based on actual calculations made by the MULTIDOC partners for their specific enterprise situation. The validation stage also included a risk analysis consisting of:

- Critical analysis of our approach.
- Analysis of changes to the human resources, the technical infrastructure and the organisational environment.
- Identification of any restrictions, constraints, risks and problems hitherto not taken into account.

In the following, we will demonstrate that the effective control of terminology helps to reduce costs at a very early stage of the documentation workflow. This is motivated by the costs that are needed to detect and repair a terminology error.

Let us assume that a unit cost of one is assigned to the effort required to detect and repair an error during the authoring stage, then the cost to detect and repair an error during the data gathering, harmonisation (synchronisation between product data and product documentation) and documentation design stages (which are similar to the requirements stages in software engineering) is between five to ten times less. Furthermore, the cost to detect and repair an error during the maintenance stage is twenty times more. The reasons for this large difference is that many of these errors are not detected until well after they have been made. This delay in error discovery means that the cost to repair includes the

cost to correct the offending error and to correct subsequent investments in the error. These investments include rework (perhaps redesign) of documentation, rewrite of related documentation, and the cost to rework or replace documentation in the field. Figure 44 below shows the cost pyramid of the different stages of error detection and correction.

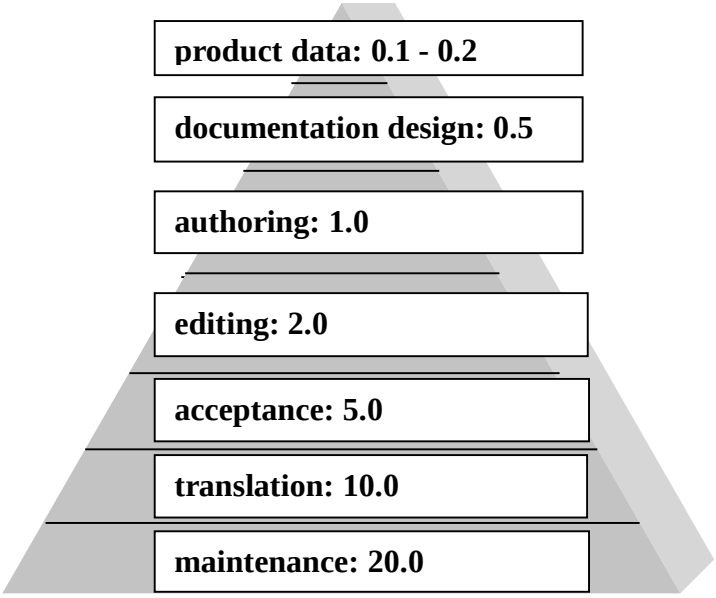

Figure 4: Cost Pyramid for Detecting and Correcting Terminology Errors

This shows that errors made at early stages in the documentation workflow are extremely expensive to repair. If such error occurred infrequently, then the contribution to the overall documentation cost would not be significant. However, terminology errors are indeed a large class of errors typically found in complex technical documentation. These errors could be between 30 % and 70 % of the errors discovered in technical documentation.

It seems reasonable to assume that a 20 % or more reduction in terminology errors can be accomplished at various levels of organisational maturity. Because of the multiplying effect, any such reduction can have a dramatic overall effect to our project's bottom line.

The user groups who are in the focus of our work are, on the one hand, the knowledge producers of documentation departments and translation departments (technical writers and translators, designers and engineers, and so forth) and the knowledge consumers in the automotive workshops (mechanics and technicians).

4 Conclusions and Perspectives

Both approaches, "bridging the gap" and ADF, are centred around the MULTIDOC terminological ontology as the primary information source. The parallel development allows for an optimal use of resources, and permits a straight forward implementation of the ADF based on already existing LT modules and components. It should be noted that the ontology centric approach has several benefits which have an direct impact on the most important business demands in multilingual documentation. Among them the most significant ones are that it

- supports harmonisation and standardisation between product data environments and

product documentation environments;

- ensures better control of information object production, translation and deployment because of clearly defined responsibilities and quality assurance measures, including a transparent workflow control;
- fosters a better integration of supplier information and subsidiary information (core business as well as associated businesses);
- abandons the need of an end-control within the information object production value cycle because of the distributed responsibilities with an integrated control functionality;
- fits with existing and emerging networked computing environments, including advanced agent technology.

More details on the project can be obtained from our web pages at URL:

http://www.iai.uni-sb.de/MULTIDOC.

5. References

Bevan, Nigel (1996): User-centred Design. Deliverable D3.1.2 v1.2 of IE 2016 project INUSE.

Bradshaw, Jeffrey M. ed. (1997): *Software Agents*. AAAI Press/MIT Press, Cambridge, Massachusetts, USA.

Church, Ken and Hovy, Eduard (1993): Good applications for crummy machine translation. *Machine Translation* 8,

Cockburn, Alistair (1997a): *Structuring Use Cases with Goals*. Web document at URL http://members.aol.com/cockburn/papers/usecases.htm.

Cockburn, Alistair (1997b): *Using "V-W" Staging to Clarify Spiral Development*. Web document at URL http://members.aol.com/cockburn/papers/ vwstage.htm

Englander, Robert (1997): *Developing Java Beans*. O'Reilly & Associates, Inc., Sepastopol, CA, USA.

Flanagan, David (1996): *Java in a Nutshell*. O'Reilly & Associates, Inc., Sepastopol, CA, USA.

Flanagan, David (1997): *Java in a Nutshell*. Second Edition covering Java 1.1, O'Reilly & Associates, Inc., Sepastopol, CA, USA.

Foyler, Martin (1997): *UML Distilled*. Addision Wesley Publishing Company, Reading, MA, USA.

Herdman, David (1997): No mirrors and magic here - You can write common code for Unix and NT. *Sunworld On-line*, April issue and May issue.

Howells, Ian (1997): Document Management Beyond 2000: More of the Same, a Strategic Tool or Spider in the Web?. In: *Document World*, Issue 1, IMC and AIIM publication, Powerhouse Solutions Ltd., Surrey, England.

Huntington, Samuel P. (1996): *The Clash of Civilizations*. Simon and Schuster, New York, USA.

Maguire, Martin (1997): RESPECT User Requirements Framework Handbook. Deliverable D5.1 of Information Engineering (IE) 2010 project RESPECT. HUSAT Research

Institute, Loughborough, UK.

Object Management Group (1995): The Common Object Request Broker: Architecture and Specification Version 2.0

Orfali, Robert et al. (1996): *The Essential Distributed Objects Survival Guide*. John Wiley & Sons, Inc., New York, NY, USA.

Orfali, Robert and Harkey, Dan (1997): *Client/Server Programming with Java and CORBA.* John Wiley & Sons, Inc., New York, NY, USA.

Stark, Heather, et al. (1996): *Ovum Evaluates: Document Management*. Ovum Ltd, London, England.

Reese, George (1997): *Database Programming with JDBC and Java*. O'Reilly & Associates, Inc., Sepastopol, CA, USA.

Schütz, Jörg (1996a): Combining Language Technology and Web Technology to Streamline an Automotive Hotline Support Service. In *Proceedings of AMTA-96*, Montreal, Canada.

Schütz, Jörg (1996b):. Network-based Machine Translation Service. In*: Proceedings of the EAMT Machine Translation Workshop*, Vienna, Austria.

Schütz, Jörg (1997): Utilizing Evaluation in Networked Machine Translation. In *Proceedings of TMI 97*, Santa Fe, New Mexico, USA.

Sowa, John F. (1984): *Conceptual Structures: Information Processing in Mind and Machine*. Addision Wesley Publishing Company, Reading, MA, USA.

Sowa, John F. (to appear): *Knowledge Representation: Logical, Philosophical and Computational Foundation.* To be published by PWS Publishing Company.

Tucker, Hugh A. (1996): Harmonization of SGML and STEP. ISO (TC 184/SC4/WG3/T14 Product Documentation) STEP/SGML Tutorial.

UML (1997): Unified Modelling Language (UML) Documentation, Version 1.0. Rational Software Corporation, Santa Clara, California, USA. Available on the Web at URL http://www.rational.com.

Unicode (1996): *The Unicode Standard: World-wide Character Encoding*. Addison- Wesley Publishing Company, Reading, MA, USA.