Multi-purpose vs. Task-specific Application: Diagnostic Evaluation of Multilingual Language Technologies

Jörg Schütz & Rita Nübel
IAI
Martin-Luther-Straße 14
D-66111 Saarbrücken, GERMANY
[{joerg,rita}@iai.uni-sb.de]

Abstract

In this paper we report on ongoing verification and evaluation work within the MULTIDOC project. This project is situated in the field of multilingual automotive product documentation. One central task is the evaluation of existing off the shelf and research orientated language technology (LT) components for the purpose of supporting or even reorganising the documentation production chain along several diagnostic dimensions such as the process proper, and the quality and the translatability of the process' output. In this application scenario, LT components shall control and ensure that predefined criteria are applicable and measurable to the documentation end-product as well as to the information objects that form the building blocks of the end-product. A prerequisite for the evaluation process is the thorough definition of these dimensions in terms of user requirements and LT developer requirements. The output quality then is the pivot where user requirements and developer requirements meet. For this, it turned out that a so-called "braided" evaluation strategy is very well suited to include both views. This strategy is also more adequate for our industrial framework, since the ultimate goal of any system development should aim at the effective matching of developer orientated objectives and the specific needs and demands of the intended users.

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, however, 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, and a construction or development phase, the so-called MULTIDOC Project. The first phase has been finished by the end of 1997, and the second phase has started in January 1998.

Evaluation is a task and particularly a process that is maintained throughout all project phases, so that a strict user-orientedness is ensured. In the inception and elaboration phase, we assessed several LT components for their deployment in supporting and enhancing the quality of technical documentation. For this, we defined diagnostic dimensions such as the documentation process proper, and the quality and the translatability of the process' output. These diagnostic dimensions are iteratively further elaborated during the construction phase to ensure that we will have quantitatively and qualitatively measurable improvements of the documentation value chain, and to guide the further development of the LT components.

In the remainder of this paper we will describe the prerequisites and the different steps of our LT evaluation. After a brief overview of the main user requirements in MULTIDOC that we have identified within the specific domain of technical documentation of Service and Repair Methods (SRM), we elaborate our evaluation methodology and the adopted method and principles. The user requirements have primarily guided the choice of the functionalities of the LT components which will be described subsequently. In the MULTIDOC project, the purpose of evaluating LT components is not to ultimately decide which specific component should win over another. Rather, the evaluation shall result in quantitatively and qualitatively measurable improvements of the whole documentation value chain, and shall also guide the introduction of possible extensions, amendments and improvements of the LT components according to user needs and demands. The following sections are dedicated to the discussion of the MULTIDOC evaluation principles (metrics and metric value scales) and the design of the evaluation process. In the last section, we will summarise our findings and draw some further conclusions.

User Requirements Analysis

Within the MULTIDOC application scenario, we distinguish two types of users:

- Technical writers as the information producers
- Technicians and mechanics in automotive workshops as the information consumers

Both groups have different requirements on technical documentation, and in particular on the different information objects. Technical writers have to produce high-quality documents which have to obey the general principles of consistency, comprehensibility, non-ambiguity, and process orientated preciseness which all feed into

translatability. Technicians and mechanics, on the other hand, are the consumers of this information and their own performance heavily relies on the success of the technical writers in terms of these principles.
First, we will concentrate on the description of user requirements for the first group, i.e. the technical writers, along the lines of the above mentioned principles.
During the user requirements analysis, a number of application areas for the employment of LT functionalities have been identified:

- Terminology and abbreviation consistency,
- Spell and grammar checking
- Style consistency according to corporate writing guidelines (controlled language),
- Information object search and retrieval

These areas also contribute to the reusability of the information objects in terms of form (information structuring) and content (conceptually precise description of service and repair operations).
For example, if in a repair operation the mechanic has to put away a specific part component of a car before executing a certain repair step, this has to be reflected in the repair information with the right wording. This then will also result in an appropriate and correct translation of this repair operation in a foreign language even if there are cultural differences in service and repair behaviours.
Besides the above introduced principles, the employment of LT in these areas also has an impact on the time and costs. As an example, 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 1 below shows the cost pyramid of the different stages of error detection and correction.
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.

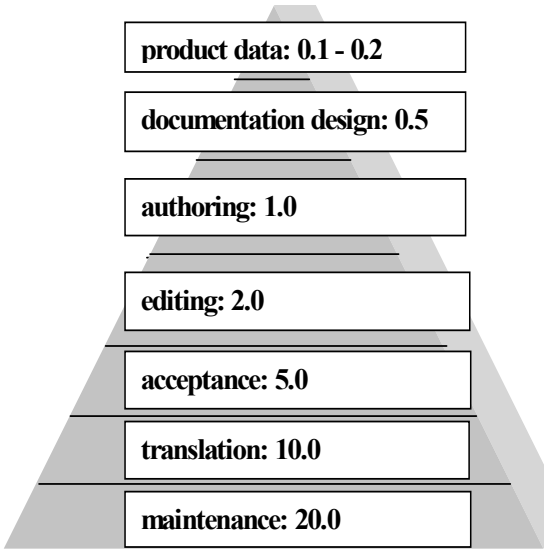

Figure 1: Cost Pyramid for Detecting and Correcting Terminology Errors

Similar calculations can be obtained for abbreviation errors, spell and grammar errors, and style errors, although their correction will be accomplished during the authoring process, i.e. the writing and composing of the information objects.
These examples profile that we are able to define measurable metrics, i.e. cost and time, for the employment of LT components which can be further classified by their contribution to the overall increase of the so-called "hit rate". The "hit rate" is concerned with the measuring of the effectiveness and efficiency of information object search and information object reusability, including the reuse of already translated information objects. This is important because today inefficient search and retrieval facilities contribute to the redundancy of information object storage, which then has an impact on unnecessary follow-up translations. Since translation in the automotive business is mostly contracted out to a translation agency, additional costs are the consequence.
Second, the information consumers in the automotive workshops need precise information in terms of form and content at the right time to assure efficient and effective service and repair measures. Here, the LT employment will contribute to certain search and retrieval operations in hotline information applications (cf. [Schütz, 1996]), including a "translation on demand" option in cases where a specific hotline information is not available in a certain language. In the latter application, the maintenance of a terminology repository that also supports domain-specific action and event readings for verbs contributes to a successful and terminologically correct "shallow translation" (indicative or informative translation) of the hotline information.

MULTIDOC LT Components

An LT component normally consists of a software part and a lingware part to which different evaluation patterns can be assigned. Whereas for the software part developers and users mostly apply the ISO 9126 "Software Quality"

standard with commercially available source code control products, such as the Logiscope system of Verilog, there is no consensus on "Lingware Quality" evaluation patterns today. The EAGLES initiative has proposed to apply ISO 9126 to Natural Language Processing (NLP) systems; however, they failed to distinguish between the two parts, and therefore we still do not have measurable metrics for lingware.

Before going into the details of our lingware evaluation patterns, we will list the LT components that we considered in our MULTIDOC evaluation work, and how the evaluation work triggered the further development of these components.

On the one hand, our goal is to support the authoring process along the above mentioned principles, and on the other hand, to foster the process of defining the form and content of information objects. For both goals the employment of LT components such as a morphological analyser and generator for a number of languages including German, English, French, Spanish, Italian, Russian and Greek, with corresponding dictionaries, including bilingual dictionaries, is in our focus.

Instead of evaluating these components as they are, i.e. with their built-in general language coverage (vocabulary and grammar), the language resources are continuously enriched with terminology, syntactic, semantic, and translation memory data.

This approach of a cyclic evaluation gave us the possibility to even apply the ISO 9126 metrics to the lingware part of the components (besides the source code control of the software part), in particular for the ISO criteria functionality, reliability, usability, efficiency, maintainability and portability, as well as the EAGLES extensions to ISO 9126 customisability and scalability. The results of the cyclic evaluations constantly feed into further refinement and improvement steps. This work also gave us new insights for the future developments of the components, especially for their deployment in networked applications as proposed in [Schütz, 1997].

Systems that can be evaluated in this way must be open, extensible and integratable on the software level through the specification of appropriate APIs, and on the lingware level through the specification of suitable "LT APIs" that permit the communication with the existing lingware resources, or through already existing system utilities that allow users to customise the lingware resources or to define their own lingware resources.

To allow for a strict user-centred evaluation process, we have also included a so-called verification step. In this step the users contribute to the finalisation of the adapted evaluation method and to the definition of the evaluation metrics. The verification process is performed on a theoretical level taking, however, into account the user's genuine working environment.

MULTIDOC Evaluation Methodology

The evaluation methodology we have adopted within the MULTIDOC project is a diagnostic evaluation. Our definition of this type of evaluation differs from the EAGLES definition ([EAGLES, 1995]) in so far as we include the user requirements of a task-specific application in our evaluation methodology. This view does not only extent the EAGLES definition, it also permits the application of the ISO 9126 quality metrics for software systems to lingware developments in a balanced way.

We call this a "braided" diagnostic evaluation. It means the systematic and regular application of predefined evaluation principles during the customisability and scalability phases of a multi-purpose LT component. These principles are then the central features of continuous quality control and progress monitoring during the evaluation process and the further development of the LT component.

In this context, the meaning of the term development is twofold. On the one hand, this concerns the software solutions of the system, and on the other hand, the lingware resources such as grammars, lexicons, translation modules, and so forth that implement the language technology proper.

The braided diagnostic evaluation methodology is defined in terms of:

- User and developer requirements which define the aimed at or needed functionality and the existing functionality of an LT component.
- Evaluation metrics and value scales for multi-purpose and task-specific applications in terms of usability (deployment potential), reliability (stability in different application scenarios, cf. above), efficiency (throughput capabilities according to time and space considerations), maintainability with respect to future customisability and scalability of the LT component.
- Process steps according to the task-specific evaluation principles consistency, comprehensibility, non-ambiguity, and operation orientated preciseness, which all contribute to the more general principle of translatability.

The actual diagnosis is then similar to a fault tracing procedure along the specifications of a symptom tree or graph. The edges of the symptom tree specify a certain phenomenon and the nodes trigger appropriate actions.

A phenomenon can be derived from the predefined principles and evaluation metrics in terms of an error classification, and an action defines a certain measure for the error repair.

For example, if a defined style checking rule does not apply according to a pre-selected set of input structures (evaluation test suites), then a possible repair operation has to further identify possible error locations as well as associated steps for finally fixing the cause of the error. Such a repair operation has to obey subsequent tests to ensure that the error fixing is monotonic.

MULTIDOC Evaluation Process

The described evaluation methodology, and in particular the actual performance of the evaluation method, turned out to be very well suited for the MULTIDOC application since the evaluation process fostered in addition the communication between users and developers, and therefore, a common understanding of the different procedures can be maintained at each stage of the project. This also minimised the risk potential of the LT developments, so that the users were not surprised about the results and possible side-effects of the LT component's behaviour.

To reach this, the following identification and specification steps are necessary for the definition of the evaluation process:

1. Identification of the task-specific application scenario, including the description of the intended user community.
2. Definition of the evaluation method and the evaluation metrics and value scales for the assessment of the quality of the output. For this, we propose task-oriented dimensions, such as the application process proper, and the quality and translatability of the process' output, which relate the developers' interests from the linguistic point of view and the economic interests of the intended user as well as the intended user community. Evaluation measures then are defined along the lines of ISO 9126 for the lingware evaluation (cf. above).
3. Specifications for the adaptation process of the existing LT component, i.e. the system "as is" including its language resources, resulting from the evaluation steps performed on a general level, especially metrics such as maintainability, customisability and scalability (cf. the discussion of APIs above). Based on the results of the evaluation of the current component, specifications can be developed which yield at the optimisation of the system's performance with respect to the predefined evaluation metrics. These specifications relate to concrete requirements resulting from the specific application domain (see above), for example, the treatment of a certain information type, typical linguistic phenomena (controlled language), use of domain-specific terminology, and so forth.
4. Aspects related to the given information technology infrastructure, for example, a network-based deployment including evaluation strategies that could be fulfilled by "intelligent" software agents (cf. [Schütz, 1997]).

As already outlined above these process steps are performed in iterative cycles. The continuous communication between the users and the developers ensures that the different evaluation patterns are applied in an optimal way, and that feedback is given on a regular basis. In addition, this processing strategy permits the adaptation or even the redefinition of evaluation patterns (metrics and value scales).

Conclusion

In this paper we have introduced the MULTIDOC evaluation methodology based on diagnostic dimensions and performed through a cyclic processing technique (method). The utilised methodology is entirely user-centred with additional support through developer orientated requirements to sanction a "braided evaluation".

This approach allowed for a clear distinction between the software level and the lingware level in the evaluation process, and the applicability of the ISO 9126 quality metrics to both levels on a thorough foundation.

The users of the MULTIDOC project agree on the fact that this approach should also be the standard approach to be adopted by LT vendors to support the integration of an LT component into an existing industrial workflow. Today, neither LT vendors nor LT OEM service providers operate in this way.

One of the future next steps is the investigation into automatisable processes to permit the development of source code control facilities for LT components, which are similar to the existing software source code control tools.

Acknowledgements

The MULTIDOC project is partly funded by the European Commission under contracts LE3-4230 and LE4-8323. The content of this paper does not reflect any official statement of the European Commission or the MULTIDOC project partners. The responsibility for the content is solely with the authors of the paper.

References

[EAGLES, 1995] EAGLES Report of the Evaluation Working Group. Geneva, Switzerland.

[Haller 1996] J. Haller: *MULTILINT - Multilingual Documentation with Linguistic Intelligence*. In: Proceedings of 'Translating and the Computer', ASLIB, London, Great Britain.

[Maas, 1998] H.D. Maas, 1998. Multilinguale Textverarbeitung mit MPRO. In: Lobin, G., Lohse, H. Piotrowski, S and Poláková, E. (Eds.): Europäische Kommunikationskybernetik heute und morgen, KoPäd, München, Germany (1998), pp. 167-173.

[Nübel, 1997] R. Nübel, 1997. End-to-End Evaluation in Verbmobil I. In: Proceedings of Machine Translation Summit VI, San Diego, California, USA, pp. 232-239.

[Schütz, 1996] J. Schütz, 1996. Combining Language Technology and Web Technology to Streamline an Automotive Hotline Support Service. In: Proceedings of AMTA 96, Montreal, Canada.

[Schütz, 1997] J. Schütz, 1997. Utilizing Evaluation in Networked Machine Translation. In: Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI) 1997, Santa Fe, New Mexico, USA, pp. 208-215.