

MULTIDOC

Authoring Aids for Multilingual Technical Documentation

Johann Haller
IAI
Martin-Luther-Str.14
D-66111 Saarbrücken
e-mail: hans@iai.uni-sb.de
internet: www.iai.uni-sb.de

1 Introduction: Technical documentation in the era of globalization

Technical documentation is getting more and more importance in the process of production, maintenance and waste disposal of goods, machines and systems. The way it is written in the original language has a strong influence in readability, understandability and translatability into a growing number of languages. Any quality enhancement in the production chain of technical documentation results in multiple profits in workshops, translation bureaus and official admission processes. This is the reason why there are worldwide efforts to use software systems with linguistic intelligence in order to help technical authors with their difficult task.

The **MULTIDOC** project focuses, therefore, on the deployment of human language technology (HLT) in SGML-based information management environments to achieve increased delivery precision, clarity and quality of the information products. The expected benefits will consist in reducing the production and translation lead times and in cutting the overall costs. In particular, this includes a business process oriented translation methodology which shall be maintained through all information production stages.

MULTIDOC is an initiative to establish a common basis for collaborative efforts of the European automotive industry in the production, management and translation of technical after-sales information. The project is partly funded by the European Commission within the Language Engineering sector of the Fourth Framework Programme. The members of the project consortium are Renault, Rolls-Royce Motor Cars, Volvo and Bertone from the automotive industry, and ITR and STAR from the translation industry; other European car makers, such as BMW (which was involved in a preceding German National project, MULTILINT), Jaguar, Rover, Saab, and so forth, are involved in the **MULTIDOC** Interest Group which is an accompanying and benchmarking body of the project.

A German version of MULTIDOC is running successfully at BMW Munich, and distributed as test version to a couple of other industrial companies; a new English version is being developed for US customers from the soft- and hardware sector.

2 Linguistic Tools

The linguistic tools which have been developed in the 15 years of IAI's work in R&D projects can be divided in the following four components:

- Source text control
- Translation aids
- Term Mining/Managing
- Content Analysis

They are preceded of, and based on an exhaustive linguistic analysis of the text. Best results are achieved if the text exists in a structured format, namely SGML or XML compatible. Word documents can be saved as HTML docs; plain ASCII is accepted equally. The text is divided into sentences and words, and a morphological and syntactic analysis is performed.

This analysis shows best results in German, followed by English and French. Other languages have been treated at least in a prototypical way; Swedish is also applied on an industrial site.

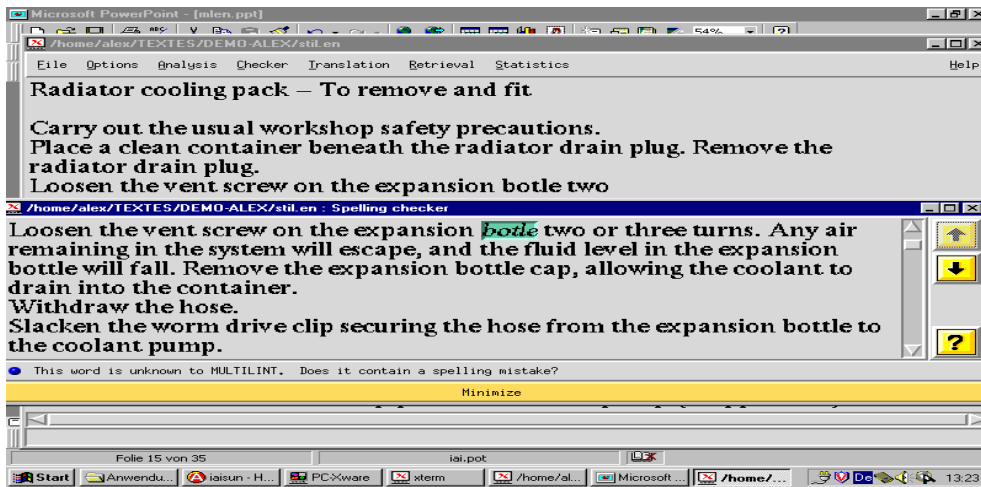
2.1 Source text control

The modules for source text control take care of the linguistic correctness of a text. This holds first for spell checking which differs in some way from the commonly known and commercially available tools. Spell checking in MULTIDOC does not use a full form lexicon, it functions on the basis of an exhaustive list of morphemes for each language. This means that every word which is built of correct morphemes and according to the morphological building rules is not marked as an error. On the other hand, possible corrections are in this system only possible for terminological entries which come from lists especially built in a certain domain.

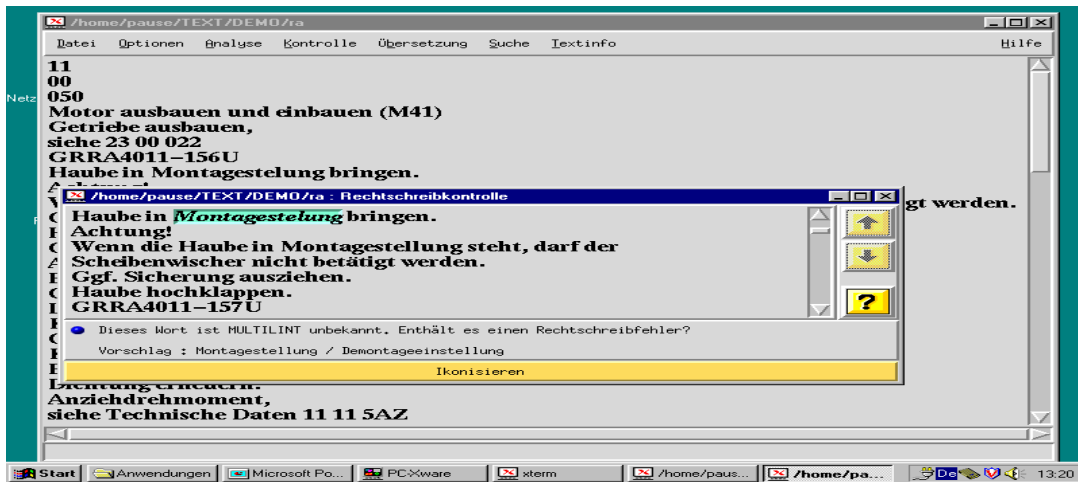
Grammar checking functions in a similar way: the system does not necessarily aim at a complete sentence analysis. It makes partial parsings of sentences and tries to identify word- or pattern-related errors which have been detected in real texts. Much material of this kind has been extracted from training manual, coming from courses for technical writing with which the technical authors have been trained at the beginning of their career. Messages can be parametrized, according to the linguistic feeling and experience of the authors, and to the respective text types.

The first function is a spell checker on the basis of an exhaustive list of morphemes of the respective language, and on the company-specific terminology. This means, that a word is recognized as correct if it is well-formed according to the morphological rule of the language concerned or if it is part of the company-specific list of terms, acronyms and abbreviations. The difference to commercially available tools consists, on the one hand, in a smaller number of marked words, and on the other hand, in more intelligent

proposals for correction. Two examples (for English and German language) may illustrate this:

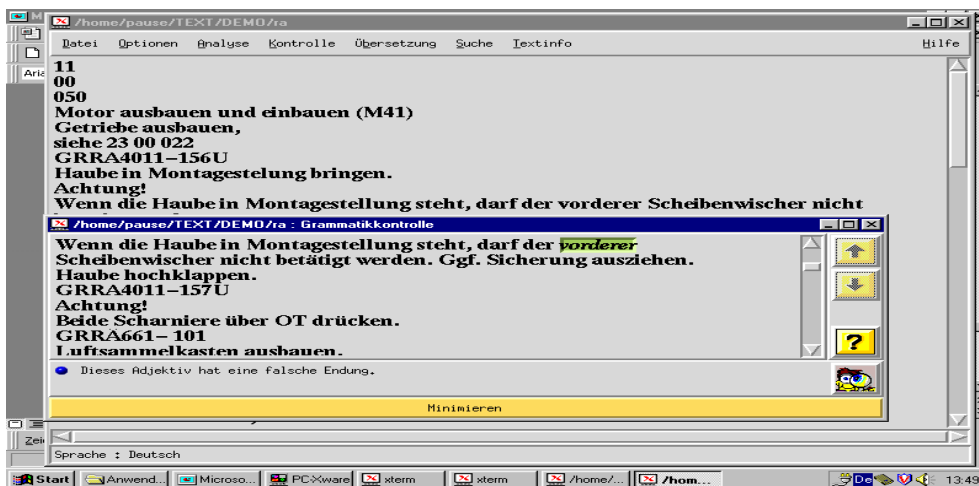


Example 1: English Spelling check



Example 2: German spell checking with intelligent corrections

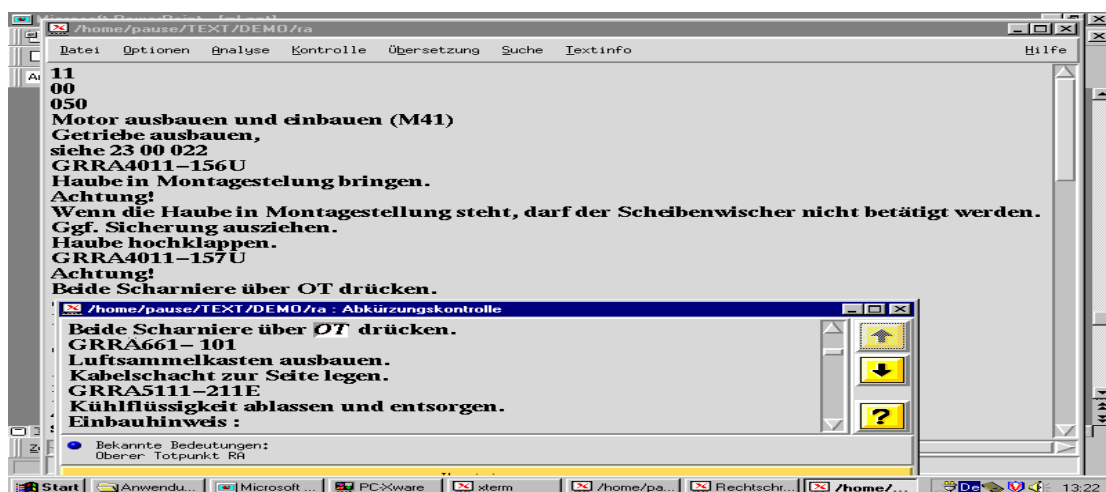
It is often the case that spelling errors result in another correct word; in this case, errors can only be detected by a syntactic analysis of the sentence, or at least parts of it. Cases like wrong agreements in German are detected in the next step which we call then Grammar Checking:



Example 3: German grammar checking

If all words are linguistically correct, the next checking step can take place. Technical Authors often tend to use and invent abbreviations because of the repetitive nature of their texts.

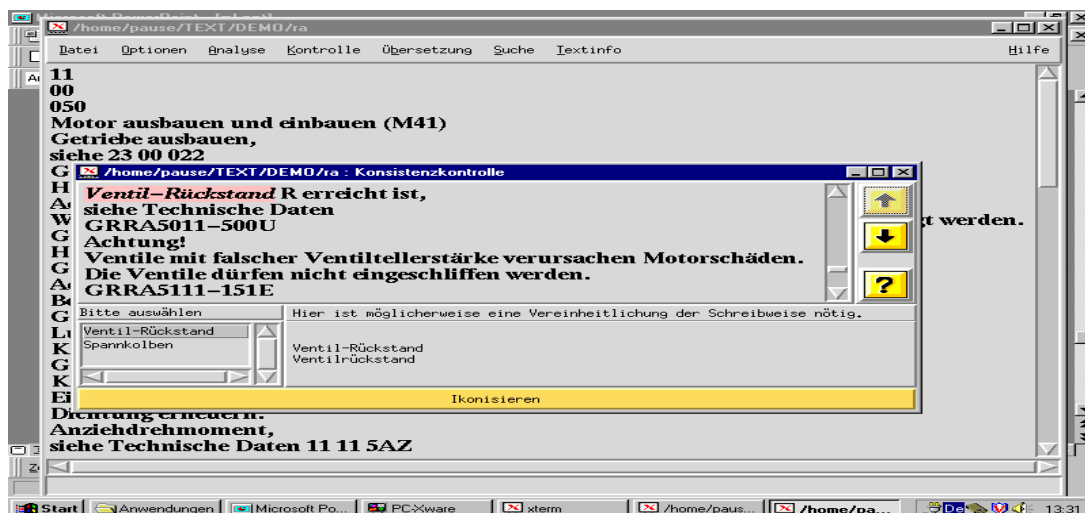
This is a correct procedure if they stick to certain principles, especially if they respect preceding uses of Acronyms. For this reason, the system remembers and marks officially approved abbreviations and their content:



Example 4: Known Abbreviations

This avoids that the author uses an approved abbreviation in a non-authorized interpretation.

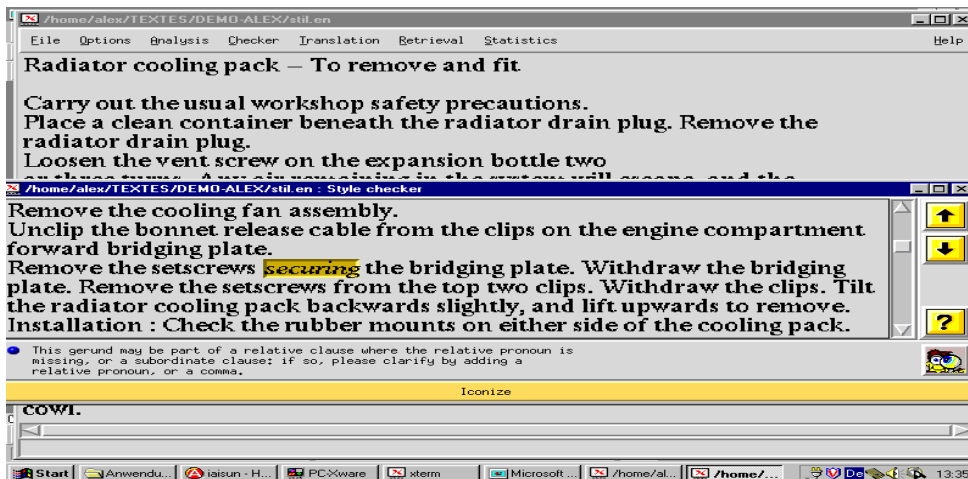
In the same way, authors tend to vary names of parts or processes, depending of the context or focus of the Repair Manual part they are working on. Using the results of the linguistic analysis as well as a company-specific term repository, a wide range of variations can be detected, from hyphenating to derivation and semantic synonymy. These variations are marked in the text, and the canonic form is displayed:



Example 5: Term variations

Correct and consistent terminology is the first basis for a company language. But several firms have created additionally so-called style guides which contain general rules for technical writing ('no sentences longer than 20 words' etc.) but also very company-specific patterns ('do never join BMW with another word, and never join it with a hyphen'. Authors are regularly trained to stick to these guidelines; in some cases, the guidelines exist even online, and certain chapters are linked to specific information elements. This is the case when the company uses already a modern information management system where small information units are created, and where authors are told not to care about formats and layout at the writing process. But experience shows that all these measures fulfil their purpose only for a restricted time, and that authors get used to have the guideline volume on their desk or on their screen. After a certain while, and when time pressure forces them to produce big volumes of documentation in short times, they do not care too much about these rules.

This is why an automatic checking tool is welcomed even by the authors themselves, not to speak by editors who have the task to control texts before they are released to the client. The goal of these actions (and also of the style control tool) is to ensure a better readable, understandable and translatable text. The latter item is often the best motivation for the use of control tools: by saving only 5% of translation time and effort, the profit is immediately multiplied by the number of languages into which the texts have to be translated. Taking an example for English, the use of 'pure gerunds' should be restricted as it can lead to misunderstandings in the interpretation of the sentence. The author will not immediately think of that possibility, as himself is aware of all underlying semantic and ontological knowledge - but this is not necessary the case with a non-informed reader.

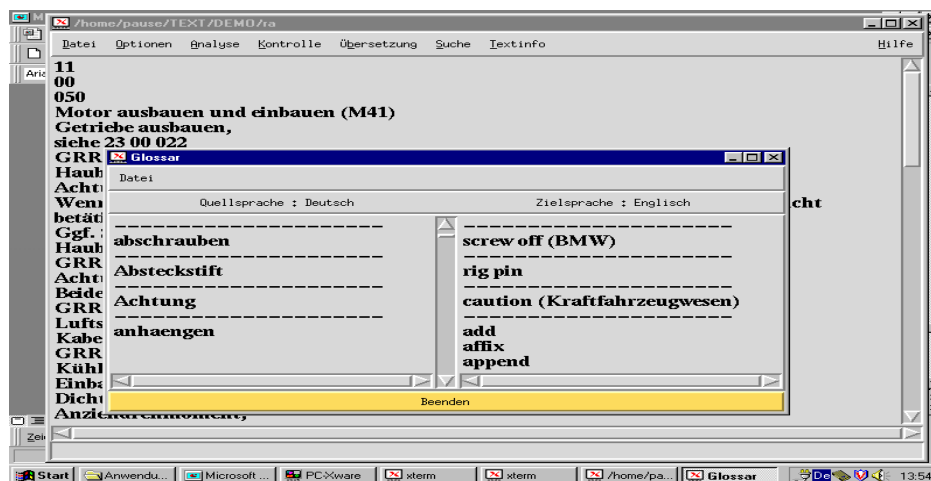


Example 6: Style checking

In the customization process, individual error messages and examples can be created for the company, the department or even the single user. The same is true for the document types where different sets of rule can apply; instructional texts undergo a much severe error search than informational texts.

2.2 Translation aids

As the linguistic analysis tools are available in several languages, a few translation aid functions have been included in the Multidoc system. The simplest form is an automatic glossary with domain markers; if a certain domain is chosen, only equivalences marked as belonging to this domain are displayed.

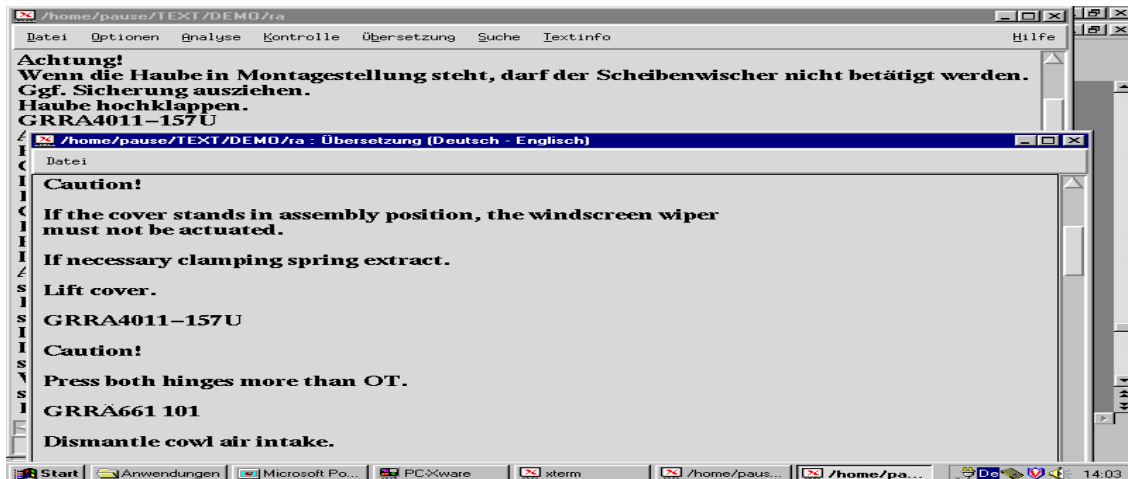


Example 7: Automatic Glossary German-English

This function is available equally by clicking on a word, as in several commercial online dictionaries. However, it includes a lemmatization of the marked word; this is especially important for morphologically rich languages as Spanish and German. In the latter case, prefixed verb forms lead to the right entry too - which is not possible in standard German dictionaries.

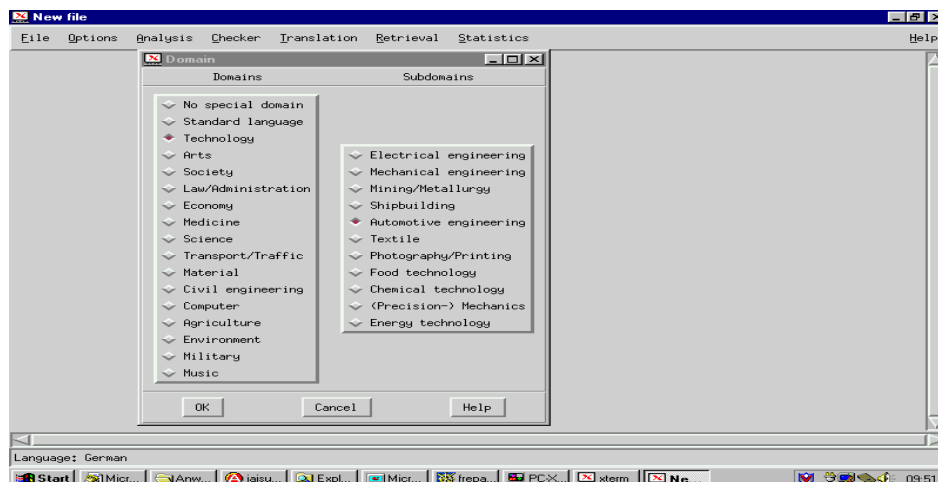
All this is a prerequisite for the tentative machine translation for information purposes. The Multidoc translation preserves normally the word order of the source language, and

only if a complete sentence analysis has been possible, an attempt is made to transform the sentence into eg. English word order. This gives a higher degree of readability to the information translation, as there are no completely deformed target sentences. This strategy has clear advantages for an information scanning and relevance checking process, but it lends to a lesser extent to the usability as a basis for human translators who are in any case sceptical in front of a machine product.



Example 8: Information translation

Whereas analysis tools are available for about 15 different languages, bilingual tools function best for couples with English and German as source or target languages. The dictionaries are divided into several domains and subdomains which can be chosen by the user.



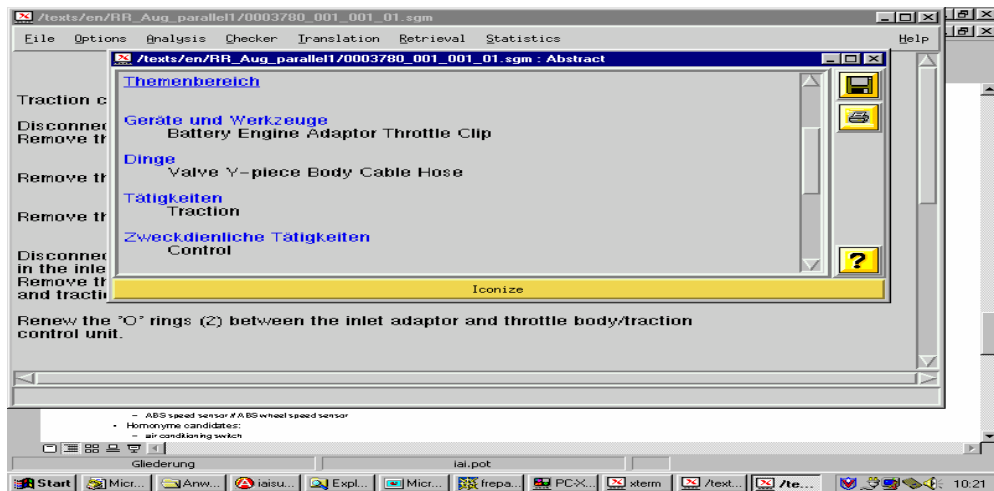
Example 9: Domains

2.3 Content functions

Another use of the linguistic analysis results is made in the retrieval component which is not string-based but indexes the lemmatized forms of nouns, adjectives and verbs. It avoids the use of truncation and wildcard strategies, and is therefore more comfortable

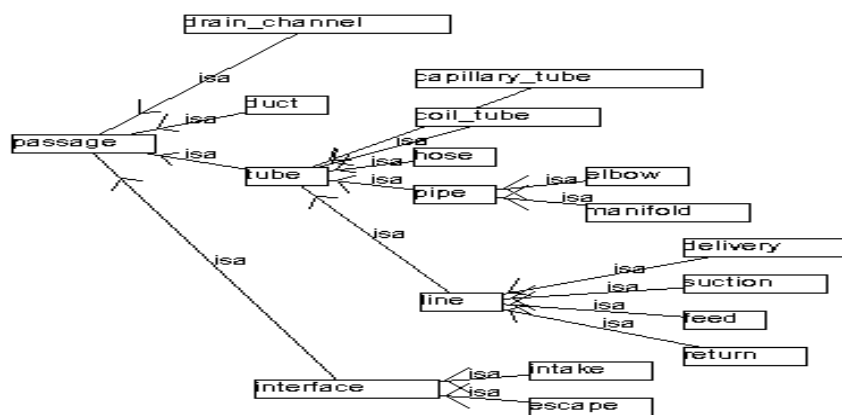
for the user without inducing the noise normally caused by string-based strategies. This retrieval can be executed both monolingually and bilingually, on the basis of the bilingual lexica shown in the glossary and translation examples.

Last but not least, some tools for content analysis and term mining complete the Multidoc toolbox. With the inclusion of statistic algorithms, a first version of an abstract can be generated automatically. A calculation of the most frequent terms, domains and semantic fields gives the user a list of the most important sentences and words from the text.



Example 10: Content analysis

After the analysis of the text corpora coming from a company, ontologies can be constructed which contain this kind of knowledge in an organized form. It is planned to use this formalized knowledge, at a later stage, for advanced controlling features. It may then be possible to control the logical sense of a repair manual, so that all steps are described in the same order as they have to be performed in reality.



Example 11: Ontology

2.4 Term Mining

On the basis of all this knowledge, it is possible to apply the linguistic and statistic algorithms to a text again, and to extract new candidates for terminological entities. The system indicates the candidate and gives the context which the specialist needs to decide if he wants to put this entry in the database or not. When parallel texts are available, the possible equivalences can come both from the bilingual dictionaries and from the corresponding sentence in the other language.

fuel recovery appliance (1)

Source: Using a *fuel recovery appliance*, drain the fuel from the tank into a sealed container.

appliance Apparat; Apparatur; Geraet; Vorrichtung

fuel recovery Brennstoffrueckgewinnung

fuel tank unit assembly (1)

Source: Remove *fuel tank unit assembly*.

assembly Baugruppe; Bauteil; Bestueckung; Montage; Zusammenbau

fuel tank Kraftstoffbehaelter; Kraftstofftank

unit Aggregat; Einheit

Example 12: Term candidate extraction from texts

Lists of term candidates, as well as unordered collections of terms which do often exist at the translation departments of companies, can be checked with respect to variants in the same way as described above:

- **Writing and morphological variants:**
 - A-pillar # a pillar #
 - A-pillar finisher # a pillar finisher #
 - A-pillars # a pillars #
- **Reductions/extensions:**
 - A/C clutch relay # A/C compressor clutch relay
 - ABS sensor cable # ABS speed sensor cable
 - ABS speed sensor # ABS wheel speed sensor
- **Homonyme candidates:**
 - air conditioning switch
 - air conditioning system
 - air conditioning unit

Example 12: Detection of term variants

3 Integration into User Environment

As all the tools are constructed in a modular way, they can be combined freely according to the user needs. This open architecture facilitates also the integration with different authoring systems based on structured editors like Framemaker or Adept. The users controls the application of the control tools by a separate window, and reads the error messages in another window. Currently, he has to make the corrections in his own editor. This is not a major drawback when he deals with shorter documents as in the case of repair documents (2-6 pages); for other applications where longer documents are the rule and not the exception, a special version is being developed where the corrections can be made in the MULTIDOC window itself. After the corrections have been finished, the correction markers are stripped off, and the text is restored in its original format to the data base.

As one can see easily from the description of the MULTIDOC tools, it is still not possible to just take them off-the shelf and use them immediately in a new environment. Term bases may have to be built, formats may be adjusted, and a smooth integration into the systems and the workflow of the company turns out necessary – but this is also true for the so-called ‘Enterprise’ versions of MT systems. Experience shows that an integration period of 2-3 years can be considered as normal, especially when the author group is not yet prepared to accept, integrate and use such a system. The best results are achieved if the authors take actively part in this process, if they formulate their own error messages, if they take part in the decision which errors should be marked, and if they feel that their documents are becoming better and better. This can be proved, for example, if the number of phone calls they receive from the translators decreases considerably – as it was the case in the pilot company. Another phenomenon is that the authors (who are not linguists at all, rather technicians and engineers) pay much more attention how to formulate their pieces of text, and that they are happy if the system does not mark anything at all in a text. Last but not least, the responsables for the translation note that the hit rate of their translation memory is going up slightly – more sentences reach the 100%, but there are also more hits with 80 % similarity: less sentences with full rate to pay, faster translation with more examples.

MULTIDOC is the successful application of human language techniques in an industrial environment.