MULTILINT

Prof. Johann Haller
IAI
Martin-Luther-Straße 14
D-66111 Saarbrücken
Email: hans@iai.uni-sb.de

MULTILINT is a research and development project, sponsored by the German Ministry of Economy. The project partners are BMW AG (Bayerische Motorenwerke) and IAI (Institute for Applied Information Sciences).

The goal of the project is to develop a system for multilingual document production and management within which linguistic components are used in the production chain as early as possible and as efficiently as possible.

1 Technical Documentation and Language Engineering

Integrating linguistic intelligence into the production process of technical documents, a considerable rate of improvement can be achieved: automatic spell, syntax and style checking of the document and checking of terminological consistency ensures a reliable information resource as well as the generation of corresponding versions in foreign languages and thus international compatibility.

Most of the MT systems available on the market, be they PC-based versions like GlobaLink, Trados-TM or the IBM-Translation Manager or midrange computer based systems like SYSTRAN, LOGOS, METAL, provide more or less sophisticated interfaces to the components supporting document production. However, the integration of such translation tools and other components into the user's environment as well as the linking of all tools as required by the user are not realized in a satisfying way by any of these systems. There exist a few efforts in this direction, such as within the ESPRIT project 'Translators Workbench' or within the EUROLANG optimizer, where at least a common user interface for all functions has been achieved. Rank Xerox is working on Compass and TANS, translation aids based on linguistic tools and architectures.

Although most of the MT systems dispose of a large amount of linguistic resources (built over years or decades), none of them uses linguistic intelligence which is available to some extent either within or outside the system, in an optimal way. Above all they do not account for the author's support, i.e. the person producing the text is neither supported in terms of an error-free document in his mother tongue nor is he supported in connecting expressions, terms etc. (e.g. synonyms, hypernyms, hyponyms, domain specific terms, ...).

The project described in this article is situated in the domain of multilingual document processing (including the domain of document translation) within the specialized field of automo-

tive technical maintenance and support services. It is based on results of more than 10 years work in multilingual language processing in the European Union (EU) carried out within programmes such as EUROTRA, MLAP and LRE. The theoretical findings of these projects have been used to build several stable research prototypes for a broad range of linguistic tasks which are now ripe to be validated in an industrial context.

2 The User and His Needs

BMW has already made some efforts in the domain of document production and document management: they developed an authoring and documentation management system RS („Redaktionssystem" - technical writing system). There are about 80 to 100 persons of the central technical service department working daily with the system producing about 600 pages per year and person. Most of the authors are engineers and/or technical writers with good knowledge of text processing and layout; however, about 25% are beginners.

The component TIS („Technisches Informationssystem" - technical information system) allows the most important service related processes to be described in a coordinated way and aims at providing the BMW dealer organization with service information by means of a CD-ROM. At the moment there are about 3500 systems installed all over the world. There exists a CD for German, English, French, Italian, Dutch, Swedish, Spanish, US English and Japanese which is up-dated four times a year.

A special application within TIS is the Hotline Support System HUS. Problems occurring with cars which are reported to BMW by appointed dealers are dealt with by means of the HUS system. On the basis of type-specific data and problems Hotline Information (HLI) already stored contributing to the solution of the problem is searched for and reported to the dealer. In case there is no HLI (hotline information) linked to a specific problem, it is checked whether the problem is already described within the HUS system, if not, the problem will be added to HUS and a corresponding full-fledged HLI will be produced and entered into the database of the system, then being available for all users.

The user needs emerging from this scenario can be characterized as follows:

In order to achieve high quality documentation in the source language, which is a prerequisite for high quality texts in the target language, the user, i.e. the technical author, needs tools allowing the text to be controlled. This control should include grammatical, stylistic and terminological checkings, thus allowing the user to process the documents optimally, to structure them and to represent their content unambiguously. This leads to an optimization of the subsequent translation process in terms of time, costs and quantity.

By integrating the MULTILINT system into the specific application environment of TIS the goal of which is having access to the above described messages in various languages within reasonable time, it will also be integrated in the overall environment of the authoring system RS, thus in addition supporting other text processing activities, guiding and advising the user through a user-friendly front-end.

Similar experiments have been carried out for English in the Aerospace industry (AECMA 95, Hoard 92, Wojcik 96); a first approach for German is desrcibed in SDD 96 and Schachtl 96.

3 Components of MULTILINT

In the following chapters, the components of MULTILINT are described in more detail; currently, they all exist in a prototypical form. Experimental use by selected technical authors at BMW is foreseen for the next months.

3.1 User Interface SYSCON

The user communicates with the tools via a common interface for which a data exchange format has been specified. The current prototypical version of the user interface has been realized in Tcl/Tk, allowing network access to the tools on UNIX workstations which are the platform used at BMW site; a port to the platform independent programming language JAVAis underway.

The user interface guides the user by suggesting the functions to be activated next on the basis of the output of an accomplished task. Every tool can also be used on its own, either by a human user or as a particular function within a technical information system (integration in a specific environment of a user).

The linguistic resources are shared among all tools in order to ensure completeness, coherence and flexibility across linguistic applications.

To ensure optimal functionality, the linguistic tools are parameterized with respect to various functions. These parameters may either be pre-defined or may be chosen by the user:

- Text format, e.g. SGML as used by the pilot user system TIS

- Source and target language, e.g. German, English, etc.

- Domain, e.g. sale, construction, logistics etc. This is organized in a hierarchical way.

- Function for bilingual or monolingual treatment, e.g.

    - spell, syntax and style checking
    - indexation (storage of an expression, e.g. abbreviations and their meaning, main topics of the document etc.)
    - mono- or bilingual retrieval
    - terminology checking
    - automatic generation of glossary and production of additional material for the use of translation memories and terminological data bases:

As mentioned, the user can choose the source and target language out of a fixed set of languages. If he does not determine the language pair, the system will set the source language which has been determined by default in his user profile. The choice of the source language is checked by means of the morphological component. If the chosen language is not compatible with the text specified, then the language which has been identified by the morphological analysis will be chosen, and this decision will be indicated. The same holds for the choice of domain.

The morphological analysis and the determination of the technical domain of every piece of text are part of the automatic pre-processing, which also provides the user with a list of words not found in the database. Commercial systems, too, have this function or a modified version

of it. Additionally, the pre-processing function described here, gives information (interactively) about the number of words which do not belong to the determined or automatically identified domain. Information about the average length of sentences is provided as well, since this is of importance for the translatability of the text.

3.2 Linguistic Resources

3.2.1 Morphological Analysis

The basic material for the morphological processing of texts already existed in various European languages: first, there is the system MPRO which decomposes the words and which provides their analysis and synthesis wrt. inflection, derivation and composition, and second, there exists a complete list of morphemes, for German, e.g. about 20.000, covering almost any text (MAAS - Mpro). These components have been evaluated with the selected domain specific texts and completed according to the needs of the specific application. For German an example of the automotive sector is given below:

```
>Seitenfenster
entry = {string='Seitenfenster',lex=fenster, clex= {lex=seite,clex=nil,
hyper= {fl=loc,f2=rand},head= {cat=n,deriv=nil,pref=nil,ehead= (gen=fem)}},
lemma=seitenfenster, hyper= {fl=loc,f2=way},head= {cat=n,deriv=nil,pref=nil,
ehead= {gen=neuter},ehead=({case=({nom};{dat};{acc}),num=sing}; {case=({nom};
{gen};{acc}),num=plu})}},graphics=cap}.[].
— Surface structure: —
[compound,n,_1619, [hyper=loc,more=way,up=an],fenster, [n, [ns,n,a,seite,
seite,_1582,_1580,_1578,(0;n),[no,1567,_1565,[hyper=loc,more=rand],[[0,0,0],
[n, 0] ], fem, no] ] ,_1531, seite] ]
— Semantic structure: —
[np, [n, [hyper=loc,more=way,up=an],_2984,fenster,_2922,_2924], [fuehren, [n,
[hyper=loc,more=way,up=an],cO,fenster,_2922,_2924],[n,[hyper=loc,_2912=
rand |_2910],wohin,seite,sg,def],no]]
=== Interpretation: ===
Ein Seiten|fenster ist ein Fenster, das zur Seite führt .
```

The concatenation of multi-word units (e.g. „in Bezug auf' - with respect to) and of terminological expressions (e.g. air intake grill) is possible. In addition, we have taken into account abbreviations and acronyms which occur quite often in these specialized texts . Also complex compounds have been accounted for in a special way.

The results provided by this component are also used for the generation of detailed concordances (e.g. extracting pieces of text showing a word of a given syntactic category).

3.2.2 Syntactic Analysis

The syntactic processing, consisting of the description of grammatically correct structures as well as the transduction rules for interpreting these rules of analysis and synthesis, already available for several languages, has been evaluated and completed on the basis of the selected corpus. Special attention has been paid to the improvement of their efficiency. In the application domain foreseen (automotive sector, especially service information), e.g. formulations in telegraphic style which are characterized by verbless constructions, subjectless verbal groups and determinerless nominal groups, have been treated.

The syntactic analysis is organized in several steps; some homographies are reduced by simple 'neighbourhood' tables and by the completion of a flat syntactic analysis which determines noun groups and some syntactic chunks of the sentence. Special analysis rules take frequent errors into account and work with relaxed constraints for the indication of syntactic

mistakes; a complete and unification based syntactic analysis is used during the attempt to obtain better translations. As this component makes frequent use of different readings and semantic constraints, such processes cannot be carried out online.

3.3 Linguistic Tools

As mentioned above, the linguistic components described are being applied for various purposes:

- for checking spelling, syntax, style and terminology
- for checking the translatability of the text
- within the proper translation process, both in analysis and in the synthesis of the target language expression

3.3.1 Source Text Control

This tool supports the author during the production of a document in the source language wrt. linguistic and terminological correctness. The source language text has to fulfil certain linguistic criteria - it would not make sense to support the production of a text version in a foreign language if it is done on the basis of a defective source language version. These criteria include orthographic, terminological and syntactic correctness, the avoidance of long sentences and of meaning ambiguities etc. They are important for a better understanding of the source language text. This type of monitoring function in text production is partially available with existing text systems which are endowed with spell checkers as well as programs checking syntax and style. The respective components in MULTILINT can however make use of an exhaustive linguistic analysis and of special features developed together with technical authors and developers of the user site systems. A few examples should illustrate the type of messages issued when inadequate input is found:

- **'Normal' spelling checking:**

      Die Käfigmutter der Sitzschiene ist lohse.

  Warning:
  „ lohse: unknown word".

- **Syntax checking:**

      Die Rohbautoleranzen sind eingehalten nicht worden.

  Warning:
  „ In this structure the negation element should occur after the finite -verb ".

      Die Änderung wird auch an der anderen Spaltdichtungen durchgeführt.

  Warning:
  „ The noun and its determiners should agree in number ".

      Es ist der zeit in Bearbeitung.

  Warning:

„A noun begins with a capital letter".

```
Die Käfigmutter der Sitzschiene sind lose.
```

Warning:

„ Subject and verb do not agree in number".

```
Dichtlippe an an der oberen Frontscheibenabdeckung krempelt ein.
```

Warning:

„ unknown structure ".

- Technical writing:

```
Kühlwasserthermostatgehalusedeckel
```

Warning:

Vermeiden Sie Wortmonster; verwenden Sie einen Bindestrich!
(Avoid long compounds)

```
In diesem Fall erhält man eine Unbedenklichkeitsbescheinigung des
Fahrzeugherstellers.
```

Warning:

Vermeiden Sie unpersönliches 'man'.

```
Eine erneute Aktivierung kann nur in der Maske 'Stimullgeräte', wie
oben beschrieben, erfolgen.
```

Warning:

Vermeiden Sie Funktionsverbgefüge.
(Avoid support verbs)

```
Die Wulstkabel , verkupferte oder vermessingte und mit Gummi umman-
telte Stahlkabel, sorgen für festen Sitz des Reifens auf der Felge.
```

Warning:

Zu viele Attribute!
(Too many attributes)

3.3.2 Foreign Language Glossary

The user may determine the lexical categories (noun, adjective, verb) for which he wants a glossary, and whether the glossary should be extended by derivational knowledge: the user may ask for instance for the translation of the corresponding verbal stem of an unknown noun.

German compounds ('Leder-lenk-rad') and multi-word units in English ('brake ventilation adapter') are looked up in the respective terminology database. If they are not found in the database they are analyzed and translated compositionally.

If it is not recommended that the machine translation of an input sequence be performed after the pre-processing or if the user does not want an automatic translation at all, he may activate the FGLOSS function instead.

Example:

```
Abdeckleiste aussen Seitenfenster hinten deformiert
```

Abdeckleiste	cover strip
aussen	outside
deformiert	destorted
hinten	rear side
Seitenfenster	side

3.3.3 Translation Memory Control - TMCON

For the rationalization of the translation task, BMW is using the translation memory system TRANSIT, a tool by means of which it is checked whether parts of the text to be translated have already been translated before. Prerequisite for this function is that there are a lot of parallelized texts available in a database.

Due to the integration of linguistic tools, as envisaged in this project, it will be possible to search e.g. for lexical elements of morphologically and/or syntactically analyzed sentences in the database. In case there are too many matches with respect to the search text, the user has the further option of a complete linguistic analysis of the text, in order to specify priorities. Within this full analysis not only lexical items but also their semantic relations within the clause are checked wrt. their similarity.

3.3.4 Informative Machine Translation - MT

Finally, there is the possibility to generate on the basis of transfer rules and the syntactic processing in the target language possible translations of short sentences or parts of texts satisfying several requirements (e.g. short messages within the service information process). The following prerequisites should be fulfilled by the input text:

- it should not contain spelling or syntax errors

- it should not be too long, respectively easy to sub-divide

- there should be sufficient lexical material in the respective dictionaries

Texts fulfilling these requirements can be analyzed, transferred and generated in the target language, based on the linguistic tools. The result will be a grammatically correct sentence in the target language which should be a good basis for understanding and, if needed, for postediting.

If a complete translation is not recommendable (either on grounds of defective input or of text complexity), a quick translation can be done on the basis of morpho-lexical information having the same word order as the input sentence or a so-called phrase translation, e.g.

```
Kühlflüssigkeit ablassen und entsorgen
```
Coolant bleed and dispose

4 Conclusion: Testing and Evaluation

During the first year of the project, the linguistic resources and tools have been adapted to and tested with the service information material coming from the industrial pilot site. At the same

time, long dialogues have been carried out between the developers and several groups of possible users of the MULTILINT system. They have expressed their wishes and preferences for their multiple purposes: technical writing of new pieces of service information in several languages, mono- and bilingual retrieval of these pieces, the control of terminological consistency for translation into new languages and the use within a future information network comprising several countries where business is growing fast. The current prototype is the result of this process and will undergo an exhaustive on-site testing during the year of 1997; in several evaluation loops, a gradual updating will take place in order to provide the best possible linguistic and ergonomic solutions for the different user groups. The project is accompanied by a working group whose members are all active in different German and other European car manufacturing companies; subject of this working group is the possibility of testing the adaptability of MULTILINT to similar purposes in the same and other domains. The project is one of several attempts currently made in Europe to bring the results of long-term research in computational linguistics and language engineering closer to industrial application needs.

References

AECMA SE-Guide (95): 'AECMA Simplified English', AECMA Document: PSC-85-16598, AECMA, Bruxelles, 1995

Fottner-Top (96), C.: 'Workshop: Erstellung von verständlicher und benutzerfreundlicher technischer Dokumentation', München: Institut für technische Literatur, 1996.

Hayes, P. J. (1994): 'Automatic Machine Translation of Technical Information without Post-Editing', Pittsburg: Carnegie Group, 1994.

Hoard et al. (92): Hoard, J.E., Wojcik, R., Holzhauser, K., 'An Automated Grammar and Style Checker for Writers of Simplified English', in: O'Brian Holt, P., Williams, N., *Computers and Writing: State of the Art,* Kluwer Academic Publishers, Dordrecht/Boston/London, 1992, 278 - 296.

Maas, D. (96): 'MPRO-Dokumentation', in: *Hausser, R. Morpholympics-Dokumentation,* Springer 1996.

Rank Xerox: 'Fact Sheet Locolex/Compass and XeLDA/TANS', Internal Document, Rank Xerox Company 1996.

Schütz, J. (96): 'Terminology and Language Engineering', *Proceedings of TKE 96,* Vienna, August 26-28, 1996 (pp. 100-109).

Schütz, J. (96): 'Combining Language Technology and Web Technology to Streamline an Automotive Hotline Support Service', *Proceedings ofAMTA 96,* Montreal, October 2-4, 1996.

SDD-Broschüre: 'Siemens Dokumentations Deutsch, Syntaxregeln für Technische Dokumentation', Internes Dokument, Siemens AG, München.

Schachtl, S. (96): 'Requirements for Controlled German in Industrial Applications', in: *Proceedings of CLAW 96,* KU Leuven, Leuven, 1996, 143 - 149.

Wojcik, R. H. & Holmback, H. (96): 'Getting a Controlled Language Off the Ground at Boeing', in: *Proceedings of CLAW 96,* KU Leuven, Leuven, 1996, 22 - 31.