FROM METAL TO T1: SYSTEMS AND COMPONENTS FOR MACHINE TRANSLATION APPLICATIONS

Ulrike Schwall and Gregor Thurmair
Gesellschaft für Multilinguale Systeme mbH
Balanstr. 57 D-81541 München Germany
ulrike.schwall@gmsmuc.de gregor.thurmair@gmsmuc.de

Abstract

This paper describes the progress which has been made to make MT systems usable in professional environments. After many years of significant investment, it was decided that the time was ripe for the METAL machine translation system to be better positioned in the market place. Two lines of action were followed:

- Introducing the system onto the PC market, using the GMS-T1 as a concrete example
- Reusing system components in customized solutions, using the AVENTINUS project as an example, which is a multilingual information processing application.

Both lines of action have far-reaching consequences for system development. But they also create new opportunities to improve the system's capabilities and flexibility.

1 Translation in the PC market

The personal translation market differs significantly from company-wide, customized solutions. In the low-end market, there are no linguistic experts around who can invest in costly training programs to master a complex MT machine — so a lower level of MT-specific expertise has to be taken into account. On the other hand, PC users have high expectations about user interfaces, text handling, and workflow support, and any PC-based solution must meet these expectations. For these user groups, lexical coverage requirements also differ from those of large companies. In a low-end PC version, for example, the lexicons should be larger, with good coverage of general vocabulary, and also more terminology in specific subject areas than is necessary for large company solutions where import of the company's own terminology is a more important issue.

What's more, due to increasing hardware and software capabilities, there is no longer a clear dividing line between low-end and high-end PC markets. Therefore, development is targeting several user groups in these different market segments. The result of this effort is a new range of PC-based applications.

1.1 T1 - the machine translation tool for the low-end market

The first product in the new range is the PC translation program T1 Standard German-English/English-German which was launched in cooperation with Langenscheidt, one of Germany's leading publishing houses, as distribution partner in June 1996.

Taken into account the user considerations mentioned above, turning METAL into a standard PC product involved more than just porting the current system to a new platform.

1.1.1 Removal of expert handling options

We considered many of METAL's fine-tuning facilities to be too complicated or even unnecessary for PC users, who must concentrate on their particular task and not on the MT tools they are using.

Therefore, several useful tools and tuning options in the high-end product were not incorporated into the standard PC version of the system. Among these are:

- Complete control of the translation process. While METAL users could interrupt the translation process at every step and check the result in intermediate files, the PC version only supports complete “logical” steps like translation, lexicon lookup, coding, and post-editing. No expert handling requiring further training is offered inside these steps. This implies that the system must be more intelligent and robust in its internal structures.
- Expert pattern matching. METAL users were able to use a pattern matcher to convert input and output strings, and prevent certain strings (like filenames) from being translated. Users could define patterns in a special language. Such a feature is very useful for professionals but not necessary to such an extent in a low-end product. We have integrated the most frequent patterns for the marking of constants in an automatic component.
- Expert lexicon coding. METAL has a special coding tool, the InterCoder, which allows extensive tuning of lexicon entries and their transfers. This tool requires special linguistic training, which is not an option in the PC market. For T1, we have created a completely new lexicon editor which accesses the underlying monolingual and transfer lexicons. The software architecture still supports the separate lexicons in the database (two transfer and two shared monolingual lexicons for one language pair). The user, however, only sees a single lexicon editor displaying the linguistic information in a user-understandable and customizable form.
- There were several other expert options in METAL such as customer-specific and product-specific subject areas and the opportunity to inspect the linguistic trees produced by analysis and generation. All these features require significant training and are not available in the first (“Standard”) version of T1.

1.1.2 Adaptation to standard PC environments

The typical low-end machine in the translation market is a 486 PC, with standard office tools, and standard interfaces. T1 is the result of re-engineering METAL to meet these requirements.

- The previous system environment was no longer applicable. No client-server solution is needed, no high-end publishing system environments (like Interleaf or FrameMaker) are to be found. These features will be provided in a professional PC version of the system. The market and users' needs define the features of a product. The first goal was the market segment of **standalone PCs.** The standalone PC user, for example, requires support for RTF documents in WinWord.
- There is no need to translate from one source into several target languages. Individual translators usually have a well-defined language repertoire, and are not normally required to translate, for example, from English into four different target languages. The standalone T1 version is modularized and offers only **pairs of languages** (English => German / German => English, English => Spanish / Spanish => English, etc.) The software and linguistic components remain flexible and can be configured for single language direction translation or for analysis only or generation only.
- Standard hardware has standard operating systems and standard programming languages. To run on PCs, the system’s software kernel was re-written in easily portable **standard environments** (like C++). As a result, the kernel became much smaller (about 60%) and **performance** increased significantly (about 3 times faster). Moreover, it became easily portable into other environments (even UNIX environments).
- To take advantage of standard software tools, the system lexicon was integrated into a standard **database.** The cost is some small data base overhead (in terms of disk space), but the benefit is robustness and backend support (e.g. if a multi-user environment is targeted). Access speed remained about the same.

- As far as application software is concerned, standard PCs have standard office environments including such programs as WinWord and other widely-used tools. It was necessary not only to integrate into PC operating systems, but also into these office tools. T1 has a completely new user interface based on the **standard PC look and feel.**
- As many PCs are now connected to the Internet, we developed a special **HTML** converter to permit translation of Internet pages from a Web browser.

As a result, the new translation system meets all of the technical and embedding criteria of a standard PC product.

1.1.3 Improved User Interface Possibilities

The next problem was meeting the requirements of text handling and user interface. While previous interfaces offered more technical and expert handling, the T1 interface uses all the features of a modern graphical user interface, including icons, Drag&Drop, and standard Windowing technology.

In addition, there are several application-specific user interface issues:

- The PC version T1 supports a **Workspace** concept: This is the defined area in which users process their documents. They import documents into the Workspace, translate them, and then export the results. As long as a document remains in the Workspace, all the links to related objects, e.g. a new word list or a Translation Memory selection list for the document, are maintained.
- The PC system supports a **Scratchpad** for fast translation of sentences, e-mail messages, and other short texts. In contrast, METAL had a "translate sentence" function mainly to test the effects of lexicon coding in a code/test/code cycle. T1's scratchpad can be used for this purpose and for fast, on-the-fly translations of inserted texts. In addition, it has a built-in editor which allows the user to write new source texts and edit the draft translation without leaving the application.
- The lexicon coding component was completely redesigned. The result is T1's **Lexicon Editor.** The goal was to make coding as easy and fast as possible for the users. Given the fact that many difficult terms which require expert coding are already in the system lexicon, and that the additional entries to be expected are rather regular in their linguistic behavior, it was possible to make the Lexicon Editor more intelligent, and much simpler to use, without losing significant quality. Special user-friendly interfaces permit the user to work in the lexicon with a minimum of knowledge and effort. Existing information can be reused, and a special "defaulter" automatically generates any additional lexicon information.
- A special **lexicon browser** was implemented for searching and inspecting the system lexicon. This was necessary because users need to know which translations the system will choose from several possibilities. If users want to change an entry in the system lexicon, they simply edit it, using the same Lexicon Editor as for new entries.
- To facilitate familiarization, **context-sensitive help** was built into the system. This feature helps the user when difficulties are encountered at specific points in translation sessions.

One should note that many of the user interface improvements require higher system intelligence than similar systems offering more user control. Simpler lexicon coding, for example, means higher linguistic intelligence is necessary for defaulting and categorizing entries.

1.1.4 Improved Workflow Support

On top of the improvements in handling and user interface, T1 provides better support of the translation workflow. Several new features have been implemented to achieve this goal.

- There are several separate stages in the translation process. Based on the workspace concept, all **pre-/post-processing** and conversion steps are carried out while a document is being translated in the workspace. All these stages are fully automatic, and intervention by the user is not necessary.
- **Integration into the word processing environment:** The standard environment for a translator is the word processor. He/she should be able to call the translation tool directly from this environment. Links to a standard office environment have been implemented, and the machine translation, Translation Memory functions, and lookup facilities are embedded into the word processing environment.
- Users can still set parameters to control how a text is actually translated. A user exercises this control via the **translation settings.** These settings allow the user to tune T1 for particular texts. These settings have been redesigned in such a way that former METAL parameters, e.g. subject area selection to control the output text quality, are still supported; but we have augmented the parameter defaulting mechanism and added some new settings; these control markups and general linguistic decisions to be taken by the translation engine. There are no parameters which require expert knowledge.
- **New Words List Editor:** The user can call up the T1 lexicon in the Lexicon Editor window, simply browse through, and make ad-hoc modifications to the lexicon at any time. However, T1 now offers a more controlled method of adding words to the system. T1 can create text-specific New Words Lists which contain the list of unknown words and compounds found by T1 in a particular document. These can be viewed and modified by the user in a specially adapted version of the Lexicon Editor window called the New Words Editor. Menu import functions or Drag&Drop facilities allow the adding of these new entries to the lexicon after modification by the user or simple defaulting by the system.
- **Dictionary Lookup facilities:** During post-editing or coding sessions, the user may want to consult a normal human-readable dictionary for information on the idiomatic use of words and phrases. To give the user the opportunity to access this type of information from within T1, lookup dictionaries, e.g. Langenscheidt's New College German Dictionary, have been integrated into the T1 applications and can be easily consulted via menu options or icons in the toolbar. An automatic **lemmatizing module** finds the base form for the lookup.
- **Background Translation** and **Translation Queue:** Once the lexicon has been updated, users can start a translation run. Documents are translated in the background and the user can follow the progress of the translation in a special Status window. Documents can be translated immediately or queued for translation at a more convenient time (e.g. during the lunch break or at night).
- **Multitasking** capability: Switching to other Windows applications is possible at any stage during a translation.
- **As** far as **translation quality** is concerned, there is no difference between T1 and METAL. Both the lexicons and the grammars have been ported without a loss in quality.
- **Quality Assurance** was a major issue in the development of the PC versions; up to 30% of the overall effort was spent on testing, workflow and user interface control, bug repair, and adaptation of the design of our translation software to meet the user's needs. Robustness was a keyword too.
- Once they have obtained the translation results, users are also supported in **post-editing.** T1 enables the user to view and edit source and target documents without leaving the application. Special color markups identify words or sentences to which T1 wants to draw the user's attention: These may be, for example, unknown words, alternative translations or a 100% perfect match from the Translation Memory database. Human-readable dictionaries, the T1 lexicon, and memory lookup facilities are easily accessible during post-editing. Post-editing directly in WinWord with all these T1 facilities is also supported.

In general, T1 supports a uniform workflow much better than METAL. It should be noted, of course, that in company-specific, high-end solutions, additional or even different workflow features may be required; there is much more variety than in a standard PC environment. However, the functions developed for the PC market are re-usable in company-specific environments as well.

1.2 Professional version

Once the standard version of the T1 program had established itself, the system was gradually upgraded into the area of professional translation by adding new features. People experienced in the translation field helped in the design. A selection of the features implemented in T1 Professional is described below:

- Integration of a **Translation Memory component** for alignment and storage of already translated documents, with translation & post-editing support. Translation Memory reduces the processing time by providing post-edited segments of text extracted from previous translation runs. The advantage of using Translation Memory in an MT context is that all resources of the MT application are available for the memory; i.e. the MT lexicon can be used in the alignment phase, leading to increased accuracy in aligning texts. In addition, a better integration of machine translation and memory translation output can be achieved, leading to easier post-editing. Integration in the T1 context means that text in a new document is first looked up in the selected Translation Memory modules. Text segments which cannot be found or where the match is not of the defined quality will be sent to the MT component for translation. Should Translation Memory contain sentences which are identical or similar to ones in the new document, their stored translations can be retrieved and written directly into a new target document, or into a special Selection List which can be processed during post-editing. The necessary degree of similarity can be freely specified by the user. The result is always a completely translated document. The percentage of Translation Memory vs. MT output will depend on the availability of memory modules for the document in question. Modules supplied with the system are Business Correspondence, Phrases and Idioms (Langenscheidt's New College German Dictionary), and Microsoft data processing terminology.
- **Memory Lookup:** In T1 Professional, it has been decided that the information in the Translation Memory should be made available to the user on demand, and not just as a fixed part of the translation process. Memory Lookup allows the user to mark a word or phrase, and then search for the selected text in one or more memory modules. The result of the search is a list of sentences in which the word/phrase occurs, together with the stored translation. This function can be an invaluable terminology aid.
- **Extended Lexicon Editor:** First of all, further modularization of the software was achieved by separating the core software from the language-specific software parts of the lexicon editor. What's more, as it can be assumed that professional translators have a better linguistic background than non-professional users, some of the options considered to be too sophisticated for standard users have been added to the professional version. These provide more linguistic control over the system's output and more powerful coding options. For example, the declination values of new noun entries are still defaulted, but these values are shown as declined noun forms in the Grammar window and can be modified by the user. Furthermore, the dependency between gender and noun inflexion values in German is taken into account in our defaulting mechanism.

  - We have added internal recognition of heads of compounds and multiword nouns; the result is displayed in a user-friendly and user-accessible/modifiable interface.
  - For German, the user has a choice of old or new spelling.
  - Display of full entry forms is a further option. Many of the entries for nouns, adjectives and verbs contain extra information, such as prepositions or object markers, which restrict their use, to specific contexts. The user can choose whether or not to display this information as part of the entry in the Browse View.

- The user can code new lexicon entries as American or British English and choose the desired "Dialect" for translations.

, - As far as verbs are concerned, we have extended the range of verb argument combinations that can be coded (e.g. two successive prepositional expressions). Further semantic tests (human/non-human) have been implemented to restrict a certain translation to a specific use of the verb. Additionally, the user is given the possibility of defining a one-to-one mapping between the source and target verb arguments by entering these in an explicit order, e.g. *sich unterhalten über etw mit jmdm -> discuss sth with sb.* To speed up coding of verb arguments, the right mouse button allows fast access to a list of supported verb frame patterns.

- **Lexicon History:** Whenever lexicon entries are added, modified, deleted, undeleted or integrated, these actions are recorded in a special Lexicon History window.
- Better system parameter tuning. For example, users are able to customize the subject area hierarchy according to their needs. This provides enhanced transfer coding possibilities. In professional translation, **subject area tuning** is a necessary feature.
- Opportunity to **import and export terminology and Translation Memory modules:** Professional translators often have card files or data bases of terminology. The MT system must support importing these into the MT lexicon. This is not just to preserve the investment that the translator has already made. It is also necessary to guarantee consistency in translation, both with previous versions of the same text and with texts that have already been translated without using MT tools. Import and export facilities are also supported for exchange of memory modules. These import and export facilities simulate the functionality of a multi-user system.

In summary, upgrading a low-end MT tool for professional use means keeping the basic features of the PC version and integrating the feature extensions necessary to support professional translators, maintaining and/or augmenting user-oriented control mechanisms, user-friendly interfaces and good workflow support. Only a product of this kind can lead to the desired productivity increase.

1.3 Further developments

T1 is currently available in a Standard version and in a Plus version that augments the Standard version with human readable dictionaries. We have now introduced a Professional version as described in the previous section. We will, of course, continue to improve T1's functionality on the PC platform and plan to add additional language pairs. But we also see opportunities for use in customized solutions, incorporating the basic MT components of T1 as part of a larger problem solution. And, vice versa, developments made in the area of customized solutions, particularly in high-end, high-volume, distributed machine translation servers, can be utilized in future T1 systems.

In cooperation with Lernout & Hauspie Speech Products, we have now started design and development of an Internet-based client/server machine translation system operating on the basis of our GMS MT technology. We had already gained experience in this area, working together with several user groups, in the course of various METAL projects [Schw95]. All language pair development carried out for this online application can also be used in T1 standalone PC versions.

We are also working on interfacing Lernout & Hauspie speech products with our machine translation technology. A first prototype was shown at CeBIT 1997.

All product development activities are derived from a single mainstream core technology line.

T1 is an example of a PC user interface driving an embedded MT engine. The MT engine is quite separate from the interface and can be easily extracted and applied in other areas. Furthermore, each component of the MT engine, the lexicons, the grammars, the parser, the translation memories, etc. are individual components which can be used in any combination to augment other programs. The GMS MT engine is an invaluable linguistic resource and offers a flexible basis for supporting non-MT applications.

The next chapter provides examples of how non-MT applications can benefit from the technology GMS has developed for T1.

2 Machine translation as a component in larger systems

The second line of action that makes MT more professional is its embedding in larger systems. There are several such approaches, e.g. to lower turn-around times in translating software bug reports [Gra95]. The example presented here is Multilingual Information Retrieval in the context of the AVENTINUS project.

2.1 The AVENTINUS context

This project [TB97] aims at supporting multilingual communication and co-operation in international drug enforcement; it combines techniques of text understanding (information extraction, intelligent indexing) with techniques of searching in structured and textual databases. Text types are open sources (mainly news agency messages) as well as internal sources (for example police reports).

In both situations, translation is a prerequisite to make a foreign language text understandable for the potential users. The requirement for translation is just that the relevance of the content of a text should be evaluated; in case it is relevant, a good quality translation follows. So the task is to free translators from the need to translate material which turns out to be irrelevant after it has been translated.

2.2 Translation technology in AVENTINUS

2.2.1 Technologies

AVENTINUS combines several translation tools to achieve this goal:

- **term substitution** is the simplest technology; it consists in inserting native language terms into foreign language text. The technology to be applied consists in basic linguistic processing (tokenizing, lemmatizing, tagging); then term candidates (which can be single words and multiwords) are looked up in the lexicon; in case of ambiguities, there is a filtering phase to remove irrelevant transfers. The resulting equivalents are inserted into the text. So users can evaluate if the content of the text is relevant or not.

Term substitution can be used in cases where no other translation means (like machine translation) are available, like cases where texts in Arabic, Urdu, Farsi etc. must be looked at.

- **Translation Memory** technology is successful in cases where texts must be considered which are rather homogeneous and repetitive. In the AVENTINUS context, this holds for police reports and internal communication. This communication is rather structured and repetitive; however, it deals with variations of certain variables, *like <person> was arrested <at place> <at time>. Do you have any information about <person>?* This type of structures can be analyzed with fuzzy matching techniques; however, as named entity recognition is one of the components of AVENTINUS, a variable match, the variables being named entities, will result in better precision of these translations.

So the Translation Memory will be enriched by variable match capabilities.

- Finally, **machine translation** will be applied to all text types which match the domain as well as the language combination in question.

In order to make machine translation work, several requirements must be fulfilled.

2.2.2 Workflow

AVENTINUS uses translation technology at three levels in the workflow:

- **If a new text** enters the system, it must be evaluated for relevance, and if relevant, be routed to the responsible person or department. This routing task can sometimes be done fully automatic, but if it is subject to human evaluation, human evaluators must be in a position to roughly understand the context of this text. In most cases this text will be foreign language material (e.g. there are just four official Interpol languages; German or Italian police officers will *always* receive a foreign language document from Interpol); so translation is a first step to make the text understandable.

This problem can be tackled by any of the technologies just described; evaluators should have the choice to select the tool they want to use.

- If an analyst has a **search** problem, he/she will have to search foreign language databases; be it structured (cf. different translations for places, different transliterations for names, etc.) or textual. Often, the optimal search term in the foreign language is not known (e.g. a German analyst searches for *Kokainhändler* and translates *cocaine dealer* which is a correct translation but the better term would have been *candyman;* so search results will be poor). So the native language search request must be translated before it can be processed.

Query translation usually has formal statements (some SQL or Boolean structures) as input; these structures must be preserved in order to have them processed correctly, and the terms in them must be translated and re-inserted. Support is needed when the foreign language is not understood by the analyst, so he/she cannot check if a translation is good or not.

- In case the search is successful then the **retrieval result** will consist of a set of hits (textual and structured). These hits may be in foreign language. Again they must be retranslated into the analysts' native language. Otherwise multilingual retrieval is incomplete. Again, all translation tools mentioned above should be available for this purpose.

In either situation, translation is the *purpose* of using the translation tools. It is always a necessary step in information processing, i.e. evaluating the content of information items.

This fact puts some constraints on the tools to be used; this also holds for the machine translation component.

2.3 Machine Translation in AVENTINUS

It is a known fact that MT is the more successful the better it can be tuned towards the domain in which it is supposed to operate. Tuning implies several aspects:

- the structure of the domain must be modelled
- the linguistic resources must be tuned
- the text types to be supported must be analyzed
- the workflow and integration must be dealt with.

2.3.1 Domain

In the case of AVENTINUS, there may be a hierarchical domain structure like ORGANISED_CRIME which in turn dominates nodes like DRUGS and CARS and TERRORISM, while DRUGS may be further divided for example according to their composition (chemical (like *Ecstasy),* or plants *(Cannabis)).* Each node will have terminology attached.

The MT system must be able to model this environment, by allowing for user-definable domain hierarchies. While several MT tools only have a fixed pre-defined topic hierarchy, the T1 system allows for such tuning.

2.3.2 Terminology

Tuning the linguistic resources mainly involves tuning the terminology. AVENTINUS will have special drug-related terminology in several languages, including definitions (which in the case of drugs is their chemical composition), relations like *slang_name_for, weakly_related, broader_term, narrower_term* etc. This lexicon is a resource which is accessed by all AVENTINUS tools, not just the MT component. It is described in more detail in [Thu97].

It is important to ensure terminological consistency between all the tools of the system; term substitution (accessing the general AVENTINUS lexicon) must produce results which are compatible with the machine translation (accessing the specific MT lexicon). As it is impossible to ask users to maintain two lexical resources, it must be possible to download and upload data to provide automatic synchronization of the lexicons.

For the MT system, this means that it must support terminology import. While the simple case (importing a nicely attributed single word entry of the term base into the MT lexicon) can be processed in a rather straightforward matter, in practice we have to worry about two phenomena:

- many of the terminological entries are *multiword* entries, with some internal inflection. So the import component has to build up a complete operational multiword representation for such an entry.
- In most cases the terminological entry will be *underspecified* from a linguistic (and MT) point of view. So additional information is needed in order to make an entry operational. As in the case of a low-end translation product, users should not be forced to enter complex linguistic information for an entry; this would reduce user acceptance drastically.

Both cases require a sophisticated terminology import function for the MT component in AVENTINUS. It must be able to handle the basic structures of terminology entries (which are mostly multiword entries), and it must apply linguistic intelligence to default the linguistic features and convert a terminological type of entry into a formal linguistic type of entry.

The interchange format for such a component could be MARTIF [ISO 12620], enriched by some features for lexicographical descriptions. In the present case, however, a direct conversion from the AVENTINUS lexicon into the MT lexicon will be preferred.

The opposite option, uploading MT entries into the AVENTINUS lexical / terminological database, must also be supported. The MT component needs a function to select entries which have been added, e.g. after a certain date, and upload them to the central lexical database.

2.3.3 Interaction

Interaction mainly refers to the data to be processed.

In AVENTINUS, data range from completely unstructured (like newspaper texts) to highly structured (like police reports) items, formats being mainly Ascii but also RTF and HTML files. Moreover, the different tools communicate via the text handling format, i.e. they read and write SGML markups in the text; e.g. if named entity recognition detects a person name or a means of transportation, the respective entity is marked up.

While each type of text (police reports, email messages etc.) may require its respective format parser, the system kernel should not be affected by these external formats. Therefore a general processing format for all AVENTINUS components has been defined (called THI, text handling interchange format), consisting of a set of SGML markups.

Of course, the AVENTINUS components must be able to understand and use this format. This also holds for the MT component; in particular, inner-text markups (like *fonts, data* but also literals of different types like *dates, persons* need to be processed). If the MT system cannot cooperate with the rest of the system (and the external applications) via the text handling format, the workflow will break down.

In the case of AVENTINUS, converters have been defined which allow the MT component to process the THI files produced by the AVENTINUS components, without loss of information. The input to the MT component is cleansed of all MT-irrelevant markups, and the result of the MT is added as a language variant to the respective text portion. The MT component is thus an encapsulated function which reads and writes proper THI constructs.

2.3.4 Workflow and User Interface

The MT component must be fully embedded in the workflow, as described above. It is important to stay in the same system environment when calling the MT tool. So easy-to-use user interfaces must be offered.

The design of such user interfaces follows the design of the host interfaces. MT tools will be called via simple buttons; parameters setting and tuning should remain outside the standard interface and be available as a special option for users who need more tuning. The focus of the user task is information evaluation, not translation; this fact needs to be mirrored by the user interface.

As in the low-end system, a special user interface for MT call-up is integrated in the AVENTINUS workflow. The MT component will be launched by this interface, and the results will be presented in the users' AVENTINUS environment.

2.4 Benefits for product development

Although it seems that professionalizing the MT system requires two lines of action which appear to be sufficiently distinct from each other, there are several areas where developments in linguistics and software engineering support the progress of both lines. The principle is to implement more intelligence in the MT system in order to make it easier to handle. There are some examples of how this could work:

- Building terminology from corpora. Tools for terminology extraction are used in cases like AVENTINUS (drug domain), where consistent terminology is not yet available. These tools can be adapted to other environments too: Although this is not relevant a priori in the standard low-end version, it can be offered as a productivity tool to professional MT users, to speed up the process of building lexical resources.
- Automatic recognition of a text language is a well established technology. Integrated into a low-end tool, user interfaces can be made somewhat easier. Attempts to translate texts from English into English (simply because users did not set the language parameter of their word processor correctly) will also be avoided.
- The same holds for subject area recognition and document routing. Again, tools developed in information retrieval contexts can be used to improve and simplify user interfaces.
- Named entity recognition will, when integrated into a standard MT environment, improve translation quality. Some of these entities, like dates, names etc. require special translations which can be taken care of safely by some special devices; these devices must be well integrated into the general translation strategies.
- Meanings of terms, instead of canonical forms, can be used for better transfer selection. The contexts of terms can be used to disambiguate the meaning (cf. [Sal91]), and word sense disambiguation technology (cf. [Ch93]) can be applied. Instead of users having to code complex transfer operations, the system would try to identify meaning variants based on the contexts of terms. This would again speed up the coding process.
- Links between terms, as produced in Intelligent Indexing, can also be exploited for a better internal structuring of the lexicon, e.g. in the case of translation variants (synonym recognition, term hierarchy

classification). Translation can be concept based, instead of word based, in larger application fields, without being forced to build up a concept hierarchy in each particular case.

3 Conclusion

PC product development has forced our translation technology to adapt to standard environments, to increase performance, to provide more user-friendly, intuitive and customizable workflow support, to increase robustness, to augment quality, to extend functionality (e.g. Translation Memory), and finally to achieve modularity which will allow us faster development due to higher reusability and higher improvability. The fertile cooperation between a strictly user-oriented product development and project-specific R&D activities makes high-end translation tools more powerful and allows the creation of low-end translation tool components which are more user-friendly and easier to handle. Our goal is to keep an even balance between both lines of action. Projects are chosen in areas where our product development needs further enhancements which can only be achieved in the longer term. These are strategic market-driven decisions. The R&D activities, such as AVENTINUS, benefit from being able to use components from our product kernel software as a basis, this in turn guarantees that the results of R&D can be reflected in our product development line.

As in other technological contexts, integrating more intelligence into a translation tool only means progress from the users point of view if it makes the tool simpler and easier to use.

References

[Cha93]: Charniak, E.: Statistical Language Learning. Cambridge, MIT Press. 1993

[Gra95]: Grasmick, D.: Machine translation at SAP. Proc. MT Summit V, Luxemburg, 1995

[ISO 12620]: ISO FDIS 12620: Terminology - Computer Applications - Data Categories

[Mil90]: Miller (Ed.), G. A. 1990. WordNet: An on-line Lexical Database. Intern. Journal of Lexicography, 3(4)

[Sal92]: Salton, G.: Effective Text Understanding in Information Retrieval. In: Kuhlen, R.: Experimentelles und praktisches Information Retrieval. Konstanz. 1992

[Sch95]: Schneider, Th.: The METAL System, Status 1995. Proc MT-Summit V, Luxemburg. 1995

[Schw92]: Schwall, U.: Testing and Evaluation - A Complex Evaluation Methodology and its Practical Use in MT Product Development. SNI-Report, Munich. 1992

[Schw95]: Schwall, U.: METAL im Netz (LAN / WAN). SNI-Report, Munich. 1995

[SchwSt95]: Schwall, U., Storrer, A.: Description and Acquisition of Multiword Lexemes. In: Steffens, P. (ed.): Machine Translation and the Lexicon, Berlin - Heidelberg. 1995

[TB97]: Thurmair, Gr., Bodenkamp, St.: AVENTINUS, Supporting Multilingual Analysis. Proc. AIPA 1997

[Thu91]: Thurmair, Gr.: METAL, Machine Integrated Translation. Proc. SALT-Workshop, Manchester, UMIST 1990

[Thu97]: Thurmair, Gr.: Ein multifunktionales Lexikon. Proc. GLDV 1997, Leipzig