A Centralized Approach to Managing Multiple Lexical Resources

Susan McCormick

ABSTRACT

The rapid expansion of SAP in markets around the world has brought with it an urgent need within the company for high-quality translation that adheres to SAP-specific terminology standards. Both the trend toward outsourcing and the increased use of automatic translation tools depend critically on quick and reliable access to official company terminology. SAP is therefore implementing a strategy that will make a central terminology database easily accessible not just to all of the people who need it (internal employees, customers, consulting agencies), but to translation and terminology tools as well. This includes the MT systems, Metal and Logos, whose data will be interchangeable with SAP database data.

1 BACKGROUND

As a large and growing international software developer, SAP has translation needs that can no longer be met using traditional translation management techniques. On-screen documentation alone is currently being translated into 26 target languages from source documents generated in either German or English. To address the expanding translation volume that has come with this rapid growth, the company has adopted more flexible strategies, notably the outsourcing of translation work to qualified agencies, and the use of new technology to automate and streamline the translation process.

A critical supporting element in this approach is the generation and management of company-specific terminology. In order for SAP to successfully outsource its translation jobs, it must be able to provide translators working in remote locations with the official terminology required for a given language and domain. And introducing new technology to make the translation process more efficient will work only if the integrity of company terminology can be assured. By increasing its reliance on outsourcing and job automation, therefore, SAP has also highlighted its need for a central store of company terminology that can be quickly and easily accessed by both translators and translation tools.

2 SAPterm and STERM

In-house translators have long used the central SAP terminology database, SAPterm, to help them with their translations. At present, SAPterm contains over 70,000 entries in approximately 130 SAP-specific subject areas. Programmed in R/3, SAPterm is accessible primarily to internal staff and can be updated by users with administrator's authorization only. Subsets of the database can be extracted and made available to external translators, customers, and partners, if required.

As SAP has expanded, SAPterm's usefulness has diminished, both in the information it contains and in its functionality. To address this, a new, second-generation SAP terminology database, STERM, is under development. STERM improves on SAPterm by offering 1) more terminological information, 2) improved coverage of the languages associated with newer SAP markets, 3) transparent links to glossaries, 4) general coding guidelines for all languages, and 5) open access to all internal SAP translators for updating. STERM also requires that terms include minimal grammatical information to allow for interchangeability with other terminological/lexical databases such as MT system lexicons or Trados‘ Multiterm.

While STERM is an improvement over SAPterm, it is considered a stepping stone to a more comprehensive approach that would make SAP terminology available, probably over the Web, to anyone with a ‘need to know.’ Open access of this sort would make it possible for SAP to take greater advantage of the linguistic expertise of its consultants, customers, and external translators by having them create directly the terminology they need for their languages and subject areas.

3 MT at SAP: Metal and Logos

The Multilingual Technology Department at SAP has used MT productively since 1991. There are currently two systems that are in active use, Metal’s German-English and Logos’ English-French. To run successfully, each of these systems must have system lexicons that are current with SAP’s terminology database. Right now, the Metal German-English lexicon contains upwards of 70,000 transfer entries in 49 SAP subject areas; the Logos installation has over 43,000 lexicon entries in 45 SAP subject areas and an additional 600 SAP semantic rules.

In order to keep the MT lexicons up-to-date, translators must constantly check official SAP terminology for changes and then make the appropriate entries and edits in the MT lexicons. This is usually done via the Metal or Logos lexicon interface, both of which support an entry-by-entry user processing mode. The process is often tedious and time-consuming, pointing up the fact that *essentially the same terminology set is being coded at least three times*[2]*, on three different platforms, in three different formats.*

4 Centralizing Terminology: The OTELO Central Lexical Database

While MT has allowed the MLT Department to post measurable productivity gains, the linguistic/administrative overhead associated with maintaining each system lexicon in isolation has appeared unnecessarily high. After several years of working around the problem, SAP decided to opt for a central repository of SAP terminological data that would be compatible with both Metal and Logos, i.e., a new database and format that would allow users to code terminology just once and then exchange it easily into other formats.

To achieve this, SAP became a full partner in the OTELO project, an EU project with the aim of developing a central translator’s environment. The environment would bring together already existing technology by offering unifying standards, formats, and interfaces for NLP products like MT, TM, and TDBs.

Central to the OTELO concept is the OTELO Lexical Database, which contains entries that can be exchanged using OLIF (OTELO Lexicon Interchange Format). With OLIF, SAP will be able to create and manage terminology in a central OTELO database and easily export its data to Metal or Logos; translators will be able to convert

[2] There are other terminology databases in use at SAP for specific applications. For instance, the Asian Language group has developed its own RDB for its members in Asia; SAP also uses Trados’ Multiterm in conjunction with Translator’s Workbench.

SAPterm/STERM entries to OLIF and load them to the OTELO database where they can be used to automatically update Metal or Logos lexicons.

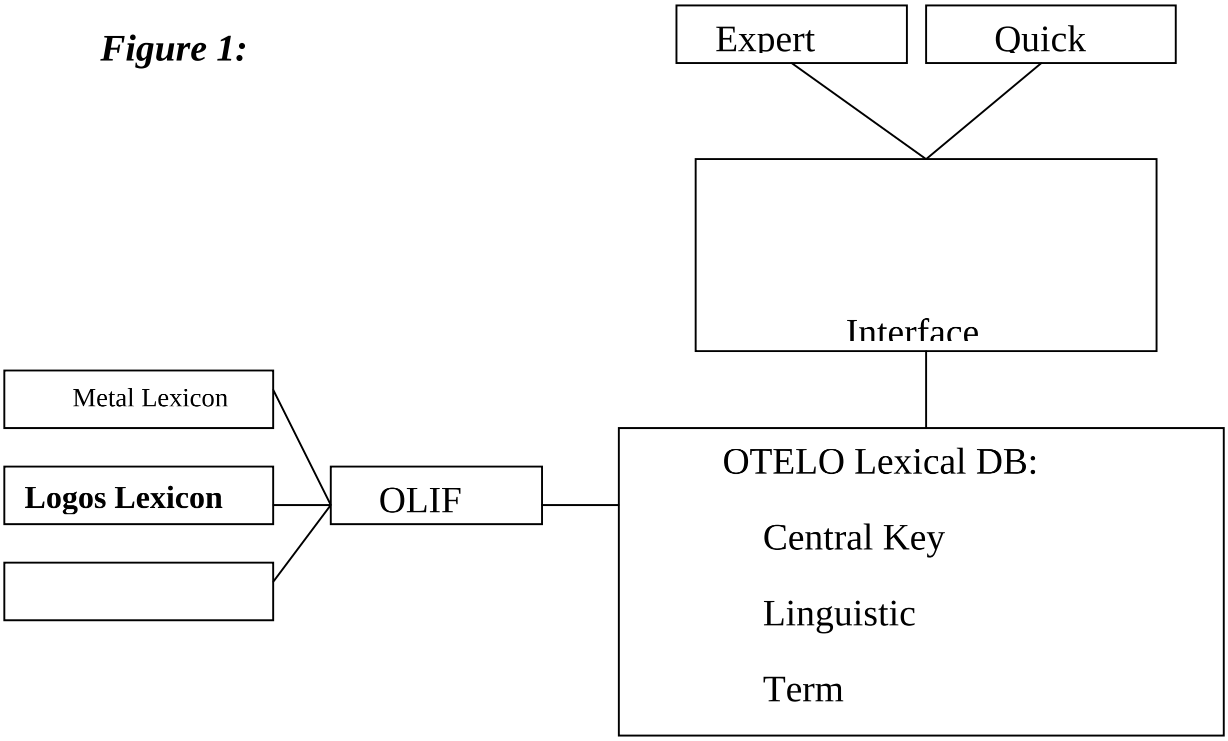
The OTELO central lexical database, designed to accommodate traditional as well as MT lexicography, has four basic partitions associated with a central key:

- Linguistic: semantic, syntactic, morphological information
- Terminology: standard terminological information, such as prose definitions, examples, user comments
- Transfer: information on transfer from language to language
- Cross-Reference: information on word relations, e.g., synonymy, taxonomy, part-whole

Using the expert interface, advanced users may code complete entries, including the specification of selectional restrictions and lexical transformations. For quick, repetitive jobs, a scaled-down quick interface is available.

With the OTELO database in place, translators in the MLT Department should, for the first time, be able to consolidate their terminology in a central location:

Figure 1:



5 Converting to OLIF

To successfully import SAPterm/STERM entries to OTELO, an interchange program must be available that takes into account some fundamental differences between the OTELO and SAP databases in terms of structure, format and defining conventions:

- SAPterm and STERM are concept-oriented; the OTELO lexical database is lemma-oriented.

- Key fields for an OTELO entry are *language, canonical form, subject, area,* and *part of speech;* the *part of speech* field is optional for SAPterm and is, in fact, rarely coded. The feature *gender* is important for MT analysis, but is also sparsely coded in SAPterm. (In STERM, both *part of speech* and *gender* are obligatory, but the current 70,000 SAPterm entries that will be migrated to STERM do not usually contain this information.)
- Conventions for formulating the canonical form are not the same. In OTELO, for example, an *adjective-noun* multiword is entered in the order that is unmarked for the given language, e.g., *adjective noun* for English or German, *noun adjective* for French. In SAPterm, conventions differ from language to language. In German, for example, *adjective-noun* multiwords are entered as *noun, adjective.* In English, they are entered as *adjective noun.* (Conventions for STERM have been drawn up to be compatible with OTELO, but, again, none of the existing SAPterm entries reflect the new conventions.)

Within the OTELO framework, SAP has written an interchange program that converts entries from SAPterm to OLIF. During the conversion process, some simple linguistic algorithms are used to automatically fill in gaps in features like *part of speech* and *gender,* as well as convert canonical form strings to the OTELO standard.

Entries can first be selectively downloaded from SAPterm so that they look a little more like a lemma-orientation than the standard SAP display. The output is a flat file of entries with a small set of lexical features:

Figure 2: Sample SAPterm Entry:

<Path>R/2: Vertrieb (RV)

<German>Abbuchen, automatisches

<Creator>FISCHERF

<Creation date>19930817

<Changed By>SATTLER

<Change Date>19930921

<Status>

<Gender>

<Category>

<Unauth.Synonym>Abbuchung, automatische

<English>automatic order filling

<Creator>FISCHERF

<Creation date>19930817

<Changed By>SATTLER

<Change Date>19931015

<Status>

<Gender>

<Category>

Figure 2 shows a case in which the German term *automatisches Abbuchen* is identified with the English term *automatic order filling* in the subject area *Vertrieb (Sales and Distribution).* Note that the feature *Category (= part of speech)* has a null value for both the German and English. The feature *Gender* is present for both terms, even though grammatical gender is not relevant to English. It has been left uncoded in the German. Also, as mentioned above, the convention for formulating the canonical form, in this case an *adjective-noun* string, is different in German than in English.

When converted to OLIF, the part-of-speech value (*CAT)* has been assigned, gender *(GD)* has been derived for German based on the morphology of the noun, the feature *gender* has been discarded in English, and the canonical forms have been regularized to adhere to the OTELO conventions:

Figure 3: OLIF Entries

<ENTRY>

<MONO>

<CAN=automatische Abbuchen>

<LG=de>

<CAT=noun>

<SA=RV>

<CE-AUTHOR=FISCHERF>

<CE-DATE=1993-17-08>

<L-AUTHOR=SATTLER>

<L-DATE=1993-21-09>

<TSTAT=new>

<GD=(n)>

<USE=online>

<ETYP=mw>

<NOSYN=automatische Abbuchung>

<SOL=R/2>

</MONO>

<XFR>

<ENTRY>

<MONO>

<CAN=automatic order filling>

<LG=en>

<CAT=noun>

<SA=RV>

<CE-AUTHOR=FISCHERF>

<CE-DATE=1993-17-08>

<L-AUTHOR=SATTLER>

<L-DATE=1993-15-10>

<TSTAT=new>

<ETYP=mw>

<SOL=R/2>

</MONO>

<XFR>

<CAN=automatische Abbuchen>

<LG=de>

<EQ=FULL>

(SA=*subject area;* TSTAT=*technical status;* ETYP=*entry type;* SOL=*industry solution*; EQ=*equivalence;* REV=*reversible*)

The SAP entries in OLIF are represented as monolingual entries (MONO), each with a full-equivalence transfer (XFR) to a target language. In addition to supplying part-of-speech information and regularizing canonical forms, the program has also analyzed the canonical forms for entry type and decided that they are multiwords. All of this derived information will be helpful not only in building the OTELO database, but also for the transition from SAPterm to STERM, since we can easily and automatically fill out information that is missing from SAPterm.

While SAPterm/STERM entries will provide the basis for new OTELO entries, Metal and Logos entries must also be converted to OLIF. For example, the Metal transfer entry in Figure 4 will be represented as well in the OTELO database:

Figure 4: Metal Transfer Entry

```
"automatische Abbuchen" NST --> "automatic order filling" NST
  Pref  S.0.0.00  Tag (SAP-RV)
<< Sap SAP Gaston 4-Feb-92 >>
```

Since the SAPterm entry in Figure 3 and and the Metal entry in Figure 4 refer to nouns with the same canonical forms, in the same languages and subject areas, the entries will be merged into a single entry in the OTELO database. During the merge, linguistic/lexical feature values will be unified where possible and the information that the entry exists both in SAPterm and in Metal will be maintained. Users will thus be able to interface with a single entry instead of managing multiple entries in several different databases.

5 Using the Central Lexical Database

It is clear that a major impetus for creating a common lexical database for SAP terminology is the need to reduce the administrative work involved in keeping several similar databases up-to-date with another. In addition to the advantages already discussed, the centralized approach should further lighten the administrative workload by offering:

- **A unitary treatment of subject-area codes:** At present, the requirements of Metal, Logos, and SAPterm mean that three separate subject-area schemas are maintained for essentially the same subject-area hierarchy. In OTELO, a single schema exists from which the others can be mapped.
- **Simple options for comparing terminology from different sources:** Merging entries in OTELO allows the user to make quick, easy checks for things like discrepant translations, i.e., cases where one system assigns one target translation for a given source word and another system assigns a different one.
- **Facility for making global changes:** Changing or deleting entries based on global criteria is easily done and applied to all relevant databases represented in OTELO.
- **Easy Import/Export of terminological data:** OLIF is an open, SGML-type format (see Thurmair et al. (1998)) to which other common formats can be easily adapted. Its coverage is relatively broad and eclectic, ranging from traditional lexicography and terminology features to the more detailed MT requirements. This alleviates the difficulties often encountered with terminology interchange.

By making it easier to generate SAP entries in different formats, the new approach should also make the startup costs for a new MT system far less onerous. With OLIF and a common lexical database, the development of a new system lexicon should require much less manpower.

In general, SAP sees the move towards integrating its various lexical and terminological information into a single, central source as a viable way of addressing its terminology needs as it expands. If translators have quick access to official terminology, if related translation tools can be brought together to support the central standard, the company will be better able to deliver consistent, clear documentation to its customers.

References

McCormick, S. (1997) "Lexical Resource Integration Requirements for OTELO," WP3.4 Technical Report, EU Project LE-2703.

Ritzke, J. (1997) “Common Lexical Resource Format/CDB Specification:CDB Features and Values,” WPA1.1 Technical Report, EU Project LE-2703.

Steffens, P., editor. (1995) *Machine Translation and the Lexicon*. Springer.

Thurmair, G., J. Ritzke, S. McCormick (1998) "The Open Lexicon Interchange Format OLIF." In *Proceedings TAMA Conference*, Vienna.