[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland; ed. Kirsten Falkedal (Geneva: ISSCO).]

The Role Of Lexicons

Susanne Schlenker
SAP AG
SAA-C METAL

Introduction of SAP:

SAP stands for Systems, Applications and Products in Data Processing. SAP develops standard software for the processing of commercial data for areas that range from accounting, materials management, to sales management. Since we have customers all over the world our documentation has to be translated into at least 8 languages (English, French, Spanish, Italian, Dutch, Swedish, Danish, and Russian).

Documentation at SAP can be divided into 4 main areas:

1. **Online help,** i.e. when someone working with SAP presses the PF1 key he finds information on the field he is currently on or on what he has to enter.
   Written with our own editor called SAPscript.

2. **Written documentation,** e.g. manuals.
   Written with WORD 5.0 + Postscript commands.

3. **Miscellaneous documentation,** including correspondence, brochures etc.
   WORD 5.0 + Postscript commands.

4. **Training material,** transparencies used in courses.
   Produced with GEM, in the future with ARTS & LETTERS.

Currently we are working with the so-called R/2 generation which runs on Mainframes and soon release 5.0 will be ready. Since this release has major changes compared to the past versions much more documentation is being written, which has to be translated into 8 different languages.

In addition to 5.0 our developers are working on R/3, the next generation, which is conceived for workstations. In this case totally new documentation will be developed which of course will also have to be translated. That adds up to a 30% rise in translation volume this year compared to last year.

Up to now the solution to the rising demand for translation at SAP has been solved by hiring more and more translators. That lead to the fact that our translation department grew enormously. In 1982 3 translators were employed and currently the department consists of approximately 75 translators, of which there are about 30 - 35 English translators (keeping in mind that SAP has about 2300 employees world-wide).

Considering the increasing translation demand, SAP has decided to turn to machine translation.

METAL at SAP

In June of 1990 Siemens installed METAL with the language pair German → English. This summer we will be the pilot customer for English → German.

Concept and organization of the project:

Since the METAL translation software runs on a Symbolics which is a standalone workstation, i.e. only one person can enter terminology into the system at a time, it is impossible for all 30 English translators to work with the system. Furthermore a relatively long period of time (approx. 6 months) is needed in order to become familiar with the system, so we would likely never go into the productive phase. At SAP the METAL group acts as a kind of service center for the rest of the translation department. Altogether three people, including myself, are working in this project. Our group prepares the texts, codes, i.e. enters the terminology defined by the translators, and delivers a rough postedited version. The translators are then responsible for the final translation.

Lexicons

When evaluating lexicons there are several aspects which have to be taken into consideration:

A. Size/Quantity

B. Quality

C. Systematically entered terminology

D. Tunability by user

A. Size/Quantity

The number of lexicon entries is by no means a sufficient criterion for evaluating MT systems. Only when taking a closer look at the inside of the lexicon can you find answers as to which system lexicon is "better". It depends on the concept of the lexicon, i.e. is it based on root forms or does it contain all conjugated and declined forms of words. If latter is so that also means more work for the user because he would have to enter more information, which is not exactly effective.

B. Quality

Closely linked to quantity is the quality of the entries.
Questions such as:

- How much information does an entry have?
- Does it contain semantic as well as syntactic information?
- Does this information apply to the source term as well as to the target term?

C. Systematically entered terminology

When assessing the contents and range of terminology some systems offer, the criteria upon which the entries are based seems to have no underlying systematic approach.

On finding entries such as "Kaiserreich Iran" – "Empire of Iran" and "Kanarischen Inseln" with adjectival and adverbial forms you quickly notice that these terms were entered for historic reasons, i.e. these words were coded at presentations or texts for potential customers.

For the user, however, it is more important that he have a basis of general vocabulary on which he can build his own lexicon.

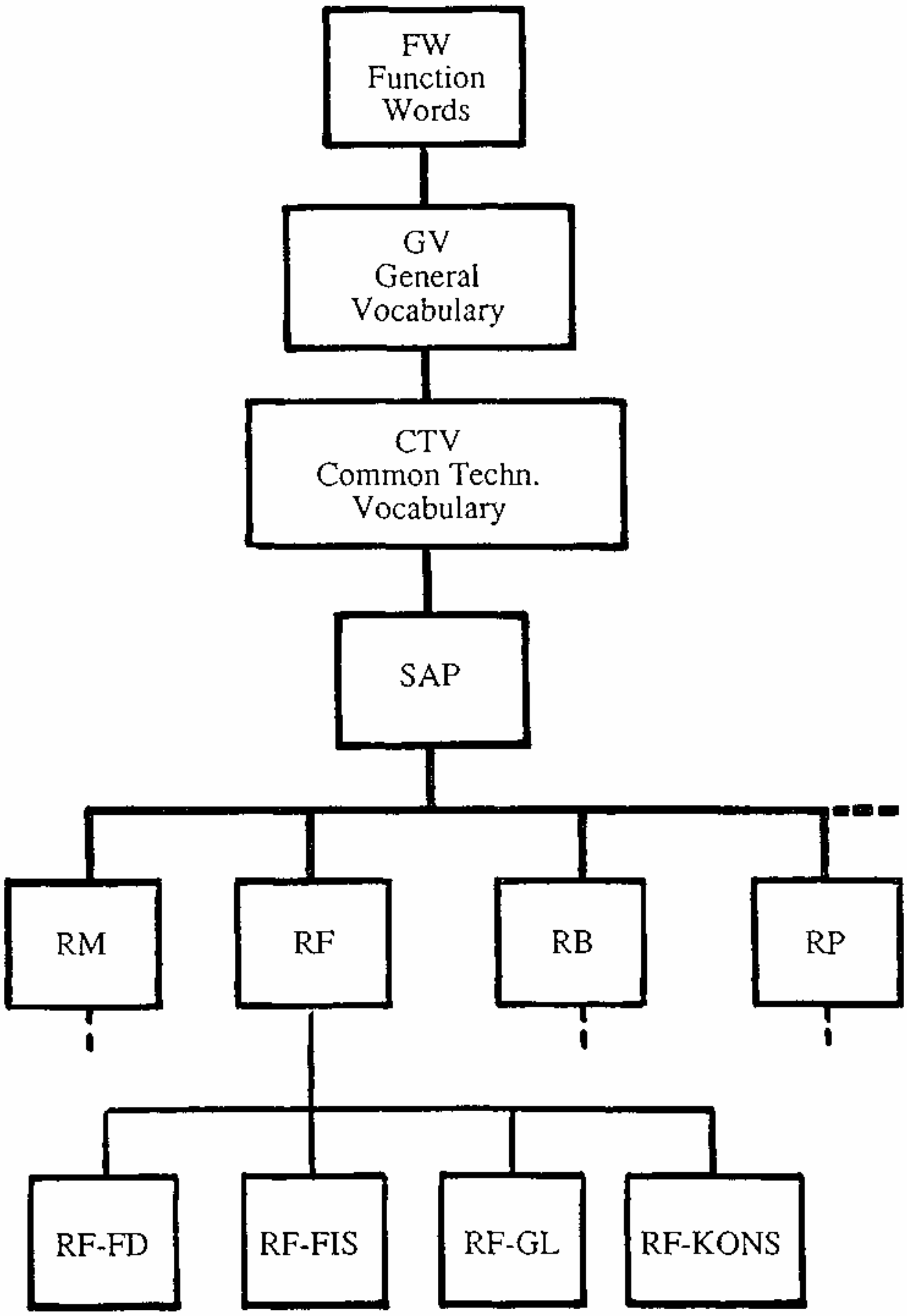
D. Tunability by user

The role of lexicons – by this I also mean any piece of information entered into the lexicon – is a very important aspect in evaluating machine translation systems. Since SAP, and probably any company, has its own specific terminology it is impossible for any MT system to deliver a lexicon containing all the words that might possibly arise in a text. Thus it is absolutely necessary for the user to be able to "tune" his lexicon himself. He himself knows best in which context the word occurs and in which cases it is translated by one term or another.

Lexicon structure

Before the actual coding can begin, however, a lexicon structure and hierarchy has to be adapted and suited to the user's individual needs. This is one possibility he has for influencing and controlling the output.

At SAP we have the following structure:



Lexicon structure and ambiguities

The lexicon structure helps deal with ambiguities that lie in the nature of language. Thus a subject code has to be assigned to every word entered into the lexicon. At SAP we have defined the following subject codes:

SAP - SAP Allgemein

SAP-QSA - Qualitätssicherung

SAP-RA - Anlagenbuchhaltung

SAP-RA-I	- Interaktives RA-Reporting
SAP-RA-OKP	- Objektorientierte Kostenplanung
SAP-RA-P	- Projekte

SAP-RB - Basis

SAP-RB-BS2000	- SIEMENS Betriebssystem
SAP-RB-EDI	- Elektronischer Datenaustausch
SAP-RB-TELE	- Telekommunikation
SAP-RB-TEXT	- Textverarbeitung

SAP-RF - Finanzbuchhaltung

SAP-RF-FD	- Finanzdisposition
SAP-RF-FIS	- Finanzinformationssystem
SAP-RF-GL	- General-Ledger
SAP-RF-KONS	- Konsolidierung

SAP-RK - Kostenrechnung

SAP-RK-A	- Auftragscontrolling
SAP-RK-E	- Ergebnisrechnung
SAP-RK-K	- Kalkulation
SAP-RK-M	- Mittelcontrolling
SAP-RK-P	- Projekte
SAP-RK-S	- Kostenstellenrechnung

SAP-RM - Materialwirtschaft

SAP-RM-CAD	- Computer Aided Design
SAP-RM-CAP	- Computer Aided Planning
SAP-RM-DIEN	- Dienstleistungen
SAP-RM-INST	- Instandhaltung
SAP-RM-KALK	- Kalkulation
SAP-RM-KAP	- Kapazitätsplanung
SAP-RM-LVS	- Lagerverwaltungssystem
SAP-RM-MAT	- Materialwirtschaft
SAP-RM-NETZ	- Netzplan
SAP-RM-PPS	- Produktionsplanung und -steuerung
SAP-RM-QSS	- Qualitätssicherung
SAP-RM-STUE	- Stücklisten

SAP-RP - Personalwirtschaft

SAP-RP-PLAN	- Personalplan

How the structure influences translation choice

Example 1:	Lieferant:	Financial Accounting	vendor
		Materials Management	supplier

If you translate a text coming from the financial accounting area and the word “Lieferant” is used, the system will automatically choose “vendor”, the correct word. This guarantees a controlled terminology choice.

You also have the possibility of having 2 or more suggestions appear in your translation.

Example 2:	Bezeichnung:	name
		description

Depending on the context which cannot be predicted before hand the translation “name” or “description” should be taken. By making 2 entries and assigning the same subject area to both, the result would be as follows:

Geben Sie die Bezeichnung des Kontos in diesem Feld ein.
→ Enter the name{description} of the account into this field.

Even though in this case the system cannot resolve the ambiguity by mere analysis the user sees the 2 options in the translation and can easily choose the required word by reading the context in which it occurs.

Coding

Since the quality of the lexicons depends to a great degree on coding I would like to give some examples of how the user himself can improve or influence translation output with METAL.

Dealing with ambiguities by using tests:

Depending on the context there are several possibilities to translate the German verb “wählen”:

1. Ich **wähle** eine Telefonnummer.
   I **dial** a telephone number.

   Test: If the direct object is “Telefonnummer”
   then use the translation “dial”.

2. Ich **wähle** die Nummer der Tabelle.
   I **select** the number of the table.

   Test: If the direct object is any word with the attribute "abstract"
   then take "select".

3. Ich **wähle** den Präsidenten.
   I **elect** the president.

   Test: If the direct object is "Präsident"
   then choose "elect"

By defining in which case the verb takes which translation the user can tune his lexicon to his exact needs.

Transformations

Example:

		DO	PP
German:	Ich ergänze	das Buch	um 2 Kapitel.
English:	I add	2 chapters	to the book.
not:	I add	the book	around 2 chapters.

With the following transformation the sentence can be translated correctly i.e. not literally:

Map	DO (das Buch)	→	PP (to the book)
Map	PP (um 2 Kapitel)	→	DO (2 Chapters)
Map	Prep. um	→	Prep. to

Whenever a DO and PP occur with the verb "ergänzen" the English Verb "add" is taken and the above mentioned transformations are carried out.

Evaluating from a user's point of view this kind of lexicon tuning is indispensable for the user in order to profit most from his system.

Other possibilities the user has to profit from his MT system are also connected to his lexicons. Since they are tailored to his needs and to his environment they contain valuable information he can use for other purposes.

He should be able to:

A. Create terminology lists according to certain criteria (all terms from a certain subject area, all nouns ...). These lists can then be used as
   - a reference dictionary for external translators (reduces in-house proof-reading time)
   - a basis for comparing different data bases.

B. Extract those files in a standard format such as ASCII.

Via files those lists as mentioned above could be compared to others – if possible even automatically. At SAP for example the translators work with a terminology data base called TERM developed by the company. Of course we want to make use of the translators' input and would like to keep that data base and the METAL lexicons consistent. It is therefore necessary to compare the two on a regular basis.

Thus MT lexicons not only serve MT systems but are also very valuable to the user for other purposes.