# When Mariko talks to Siegfried

## - Experiences from a Japanese/German Machine Translation Project -

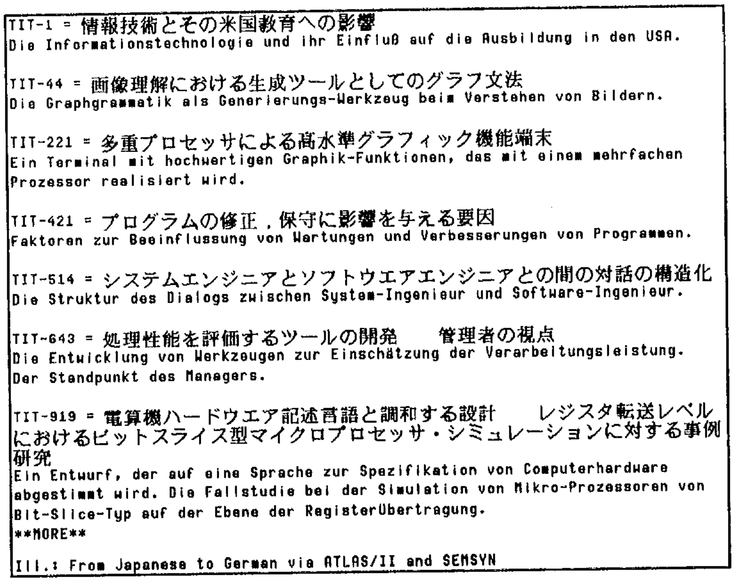
Dietmar Rösner
Projekt SEMSYN, Institut für Informatik
Universität Stuttgart,Herdweg 51
D-7000 Stuttgart 1
West Germany

### Abstract

**In this paper we will report on our experiences from a 2 1/2 year project that designed and implemented a prototypical Japanese to German translation system for titles of Japanese papers.**

### Background

An american study - published in Nature, 308 (1984) - evaluated cir. 9000 Japanese scientific papers. 75 percent of them are published exclusively in Japanese, only a 5th of Japanese papers are currently evaluated from Western refereeing and information services. The main conclusion of the study was, that the general opinion all important Japanese stuff would be published in English is not true, at least for the applied sciences. From this background and from the Japanese success in a lot of fields of modern technologies stems a wider interest in having access to Japanese material and in having help to overcome the language barrier.

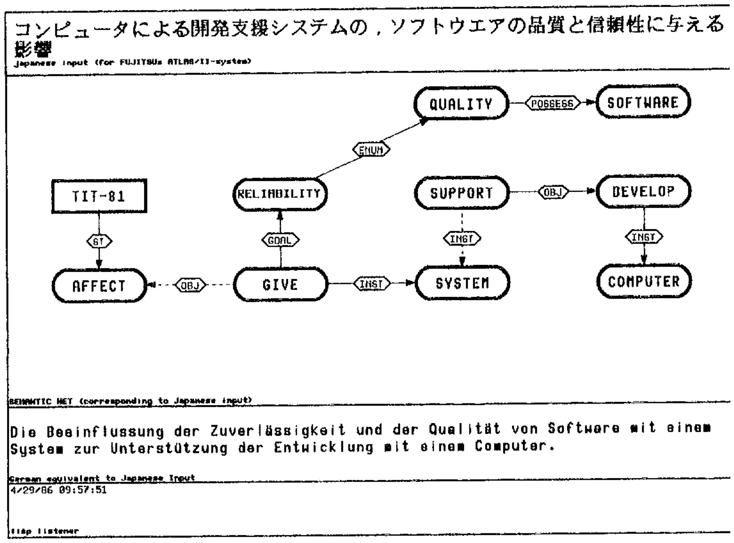


```
TIT-1 = 情報技術とその米国教育への影響
Die Informationstechnologie und ihr Einfluß auf die Ausbildung in den USA.

TIT-44 = 画像理解における生成ツールとしてのグラフ文法
Die Graphgrammatik als Generierungs-Werkzeug beim Verstehen von Bildern.

TIT-221 = 多重プロセッサによる高水準グラフィック機能端末
Ein Terminal mit hochwertigen Graphik-Funktionen, das mit einem mehrfachen
Prozessor realisiert wird.

TIT-421 = プログラムの修正，保守に影響を与える要因
Faktoren zur Beeinflussung von Wartungen und Verbesserungen von Programmen.

TIT-514 = システムエンジニアとソフトウエアエンジニアとの間の対話の構造化
Die Struktur des Dialogs zwischen System-Ingenieur und Software-Ingenieur.

TIT-643 = 処理性能を評価するツールの開発　　管理者の視点
Die Entwicklung von Werkzeugen zur Einschätzung der Verarbeitungsleistung.
Der Standpunkt des Managers.

TIT-919 = 電算機ハードウエア記述言語と調和する設計　　レジスタ転送レベル
におけるビットスライス型マイクロプロセッサ・シミュレーションに対する事例
研究
Ein Entwurf, der auf eine Sprache zur Spezifikation von Computerhardware
abgestimmt wird. Die Fallstudie bei der Simulation von Mikro-Prozessoren von
Bit-Slice-Typ auf der Ebene der RegisterÜbertragung.
**MORE**
```

Ill.: From Japanese to German via ATLAS/II and SEMSYN

### 1. SEMSYN - a Japanese/German translation system

The **project SEMSYN-83** - SEMSYN is an acronym for **SEM** antic **SYN**thesis - has produced a system for the generation of German from semantic representations. The combination of this generator with the **ATLAS/II-System** of the **Japanese cooperation partner FUJITSU** may be seen as the first Japanese to German translation system.



コンピュータによる開発支援システムの，ソフトウエアの品質と信頼性に与える影響

Japanese input (for FUJITSUs ATLAS/II-system)

SEMANTIC NET (corresponding to Japanese input)

Die Beeinflussung der Zuverlässigkeit und der Qualität von Software mit einem System zur Unterstützung der Entwicklung mit einem Computer.

German equivalent to Japanese Input

4/29/86 09:57:51

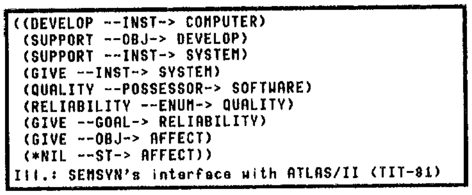
lisp listener

The analysis of the Japanese input - currently at most titles of scientific papers from the field of information technology - and its transformation into the semantic representation is the task of ATLAS/II. **SEMSYN's part is to produce a correct and understandable German text for these semantic representations**.

### 2. The overall design of the SEMSYN-System

**SEMSYN's generation** from FUJITSU's nets to German surface structures is done in three main steps.

The first step is to **transform the semantic net** delivered by FUJITSU into an expression of our own frame representation language - the so called IKBS-descriptions. IKBS stands for Instantiated Knowledge Base Schemata. This transformation does not only lead to a more structured representation, it helps as well to keep the generation modul somewhat independent from the special form of the FUJITSU interface.



```
((DEVELOP --INST-> COMPUTER)
 (SUPPORT --OBJ-> DEVELOP)
 (SUPPORT --INST-> SYSTEM)
 (GIVE --INST-> SYSTEM)
 (QUALITY --POSSESSOR-> SOFTWARE)
 (RELIABILITY --ENUM-> QUALITY)
 (GIVE --GOAL-> RELIABILITY)
 (GIVE --OBJ-> AFFECT)
 (*NIL --ST-> AFFECT))
```

Ill.: SEMSYN's interface with ATLAS/II (TIT-81)

The second – and probably most important – step is to decide **in which way the content of the semantic representation should be uttered as German text**. The output of this step is a functional description of the intended utterance in grammatical terms (IRS = Instantiated Realization Schema). The IRS description completely determines the German output. Its terminal elements are root forms of German words and their syntactic features.

```
(:NG
 (:HEAD "Beeinflussung")
 (:FEATURES (:DET DEF) (:NUM SG))
 (:POSSATTR (:NG (:HEAD (:NG-CONJUNCT (:NGS (:NG (:HEAD "Zuverlässigkeit")
                                                 (:FEATURES (:NUM SG)
                                                            (:DET DEF)))
                                            (:NG (:HEAD "Qualität")
                                                 (:FEATURES (:NUM SG)
                                                            (:DET DEF))))
                                      (:CONN "und")))
                 (:FEATURES (:DET DEF) (:NUM PL))
                 (:POSSATTR (:PG (:PREP "von")
                                 (:POBJ (:NG (:HEAD "Software")
                                             (:FEATURES (:NUM SG)
                                                        (:DET ZERO))))))))
 (:QUALIFIERS
  (:PG
   (:PREP "mit")
   (:POBJ
    (:NG
     (:HEAD "System")
     (:FEATURES (:DET INDEF) (:CAS DAT))
     (:POSSATTR
      (:PG
       (:PREP "zu")
       (:POBJ
        (:NG
         (:HEAD "Unterstützung")
         (:FEATURES (:NUM SG) (:DET DEF))
         (:POSSATTR
          (:NG
           (:HEAD "Entwicklung")
           (:FEATURES (:DET DEF) (:NUM SG))
           (:QUALIFIERS
            (:PG (:PREP "mit")
                 (:POBJ (:NG (:HEAD "Computer")
                             (:FEATURES (:DET INDEF)
                                        (:CAS DAT)))))))))))))))))
```

Ill.: IRS-Description for TIT-81

The third step – **the generator-front-end SUTRA-S** – takes the IRS description and produces a corresponding syntactically and morphologically correct German surface structure (Emele & Momma, 1985). SUTRA-S is an extended reimplementation of the program SUTRA that has been developped by Busemann in the HAM-ANS project (Busemann, 1982).

### 3. Generation from frame descriptions

### 3.1 The frame description language

The formal definition of SEMSYN's frame representation is as follows:

```
<IKBS-DESCR> :== (A <FRAME-NAME>)
                 (A <FRAME-NAME>
                           WITH . <SLOTS&FILLERS>)
                 (THE <SLOT-NAME> FROM <IKBS-DESCR>)

<SLOTS&FILLERS> :== ((<SLOT-NAME> = <IKBS-DESCR>)
                                 ...)
```

Conceptually we distinguish the following three main classes of frames:

1. **Case schemata** for verb concepts or actions (among these are all those frames that have case roles as slots).

2. **Concept schemata** for noun concepts or "picture producers".

3. **Relation schemata** – ENUMERATION, PURPOSE-, SCOPE-Relation etc.

```
(THE :OBJECT
     FROM
     (A GIVE
        WITH
        (:GOAL =
                (AN ENUMERATION
                    WITH
                    (:ARGL = (A RELIABILITY) (A QUALITY))
                    (:POSSESSOR = (A SOFTWARE))))
        (:INSTRUMENT =
                      (THE :INSTRUMENT
                           FROM
                           (A SUPPORT
                              WITH
                              (:OBJECT =
                                        (A DEVELOP
                                           WITH
                                           (:INSTRUMENT = (A COMPUTER))))
                              (:INSTRUMENT = (A SYSTEM)))))
        (:OBJECT = (AN AFFECT))))
```

Ill.: Frame-Description for TIT-81

Within this scope the repertoire of the semantic representation includes:

- "classical" case roles a la Fillmore (agent, object, method, instrument, source, goal , ...)
- roles for the further specification of actions (manner, place, time ...)
- roles for the further specification of concepts (name, concern, specialize ...)
- ways to quantify and attribute concepts
- modality (e.g. not, possible ...).
- conjunctive and disjunctive ENUMERATION.

### 3.2 Knowledge bases during generation

SEMSYN's main generation phase may be viewed as communication between two knowledge bases: **General knowledge** about principal possibilities for realizing the semantic structures – the so called **realization schemata** – and **specific knowledge** mainly about diverse possibilities for lexicalization of semantic symbols. The latter is stored within the semantic to German dictionary SLEX (Rösner, 1986).

### 3.3 Object-oriented implementation

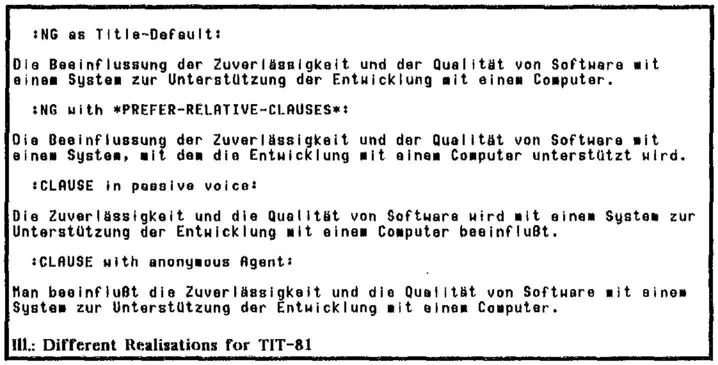
The general knowledge about possible realizations has been **implemented using the FLAVOR system of the LISP machine**. The classes of the frame representation correspond to flavor classes. Realization schemata and the knowledge about the realization of roles are defined as flavor methods. This object-oriented architecture has shown to be very flexible. It supported experimenting with the system and its step-by-step improvement.

### 3.4 Realization schemata

Frame descriptions as used in SEMSYN are recursive structures and so is – in general – the **control structure** in SEMSYN's generation. In other words: the same decisions have to be redone on each level of embedding. In embedded frames of course some decisions are already restricted by the context.

**What will be the syntactic form of the text generated for such a frame?** At least for case schemata we have as first alternative the choice between the realization types :CLAUSE and :NG (noun group). For semantic structures from titles we used as default to generate a noun group (a toplevel case schema was lexicalised as noun). Only in a few cases we had titles that had to be generated as questions like "What is a model of ...?".

If the general syntactic form has been decided upon, there are **more choices**: a clause for example could be realized as an active or a passive clause. Within a noun group the attribute could be realized as a relative clause or in the form of a prepositional group.



```
   :NG as Title-Default:

Die Beeinflussung der Zuverlässigkeit und der Qualität von Software mit
einem System zur Unterstützung der Entwicklung mit einem Computer.

   :NG with *PREFER-RELATIVE-CLAUSES*:

Die Beeinflussung der Zuverlässigkeit und der Qualität von Software mit
einem System, mit dem die Entwicklung mit einem Computer unterstützt wird.

   :CLAUSE in passive voice:

Die Zuverlässigkeit und die Qualität von Software wird mit einem System zur
Unterstützung der Entwicklung mit einem Computer beeinflußt.

   :CLAUSE with anonymous Agent:

Man beeinflußt die Zuverlässigkeit und die Qualität von Software mit einem
System zur Unterstützung der Entwicklung mit einem Computer.
```

**Ill.: Different Realisations for TIT-81**

These decisions are done with respect to several factors. One is the **type of the actually filled roles**. If a case schema for example has an :OBJECT, but no :AGENT, we prefer the passive construction in a clause realization. On the other hand **stylistic preferences** could be another factor. In the above case a preference could be to avoid passive, so we would take the realization schema "ACTIVE with an anonymus agent of 'man'".

In titles these preferences come from global switches. In real text they could come from the context.

### 3.5 Role realizations

For frames without roles - the so called **terminal structures** - the realization is more or less the lexicalisation of the semantic symbol. After this, process control and the produced IRS structure is given back to the surrounding frame or the toplevel.

If there are roles, there is some more work to be done. Some fillers of roles are **realized as distinct structures of their own** (mostly noun groups). They could be uttered for themselves.

Other roles only lead to **changes in the IRS structure of their frame**:

**-decision about syntactic features**: fillers of a :NUMBER role may e.g. lead to the pluralization of the noun group of the modified frame.

**-creation of noun compounds as head of the actual nominal group**: the filler of a :NAME role may become a prefix ("das SEMSYN-Projekt"). This holds as well for the terminal filler of a :SPECIALIZE role (variant: realization as an adjective). A negative :MODALITY could - in a noun group realization - lead to the prefix "Nicht-".

For those frames that have roles with realizations of their own this procedure recursively repeats for the frame descriptions of the fillers of those slots.

For realized role fillers it has to be decided how their IRS-structure shall be integrated in the overall structure (mostly as prepositional group) and which syntactic features could additionally be inferred.

### 4. Inferring of missing information

SEMSYN's generation modul starts from a semantic representation that was designed to be language independent. For the primitives used - especially for the semantic relations expressed by the arcs in the semantic net - this may be true.

On the other hand the data delivered to us by FUJITSU are **not really universal representations**. The fact that the semantic nets are derived from Japanese is recognizable if one looks at **the information that is not explicitly represented**.

**In Japanese number or definiteness of nouns or time of verbs normally is not expressed** - correspondingly our data do not have semantic correlates for these features (except in the rare case when they have been expressed in the Japanese original). The Japanese reader infers the missing information from the context. In titles there is no such context available. For correct and acceptable German on the other hand we need determiners and our nouns need a number. Therefore we had to develop heuristics to reconstruct this information.

**Some examples of such heuristics:**

- a nominalized case frame has to be realized with definite article in singular ("Die Generierung natürlicher Sprache").

- the :OBJECT role of a nominalized case frame should be realized indefinite and plural ("Die Generierung von Titeln"), except in cases with an exception information in SLEX ("Die Wartung von Software").

- concepts that have a :NAME role will be realized definite and singular ("Die Fourier-Transformation").

If no heuristic is applicable and if no SLEX information is found we use as title defaults 'indefinite' and 'singular' ("Ein Verfahren").

### 5. Concluding remarks

Our current concern is to **broaden the applicability of SEMSYN's generator for German**: On the one hand we are experimenting with the **generation of full texts** (e.g. newspaper stories), on the other hand we are **extending the repertoire of feasible semantic structures that may serve as input for the generator**.

**Acknowledgement:**

**SEMSYN-83** has been funded by the **West German Ministry for Research and Technology (BMFT)** from July 1983 till February 1986. Special thanks to all the colleagues that collaborated - for shorter or longer periods - within this project: Kenji Hanakata, Joachim Laubsch, Arek Lesniewki and Shoichi Yokoyama.

**References**

Busemann, S. (1982) "Probleme der automatischen Generierung deutscher Sprache", HAM-ANS Memo 8, Universität Hamburg

Emele, M. & S. Momma (1985) "SUTRA-S - Erweiterungen eines Generator-Front-Ends fuer das SEMSYN-Projekt", Studienarbeit, Inst. f. Informatik, Univ. Stuttgart

Laubsch,J., Rösner,D., Hanakata,K., Lesniewski,A. (1984) "Language Generation from Conceptual Structure: Synthesis of German in a Japanese/German MT Project", in: COLING-84, Proceedings, Stanford

Rösner, D. (1986) "SEMSYN - Wissensquellen und Strategien bei der Generierung von Deutsch aus einer semantischen Repräsentation", in: Batori & Weber (Eds.) Neue Ansätze in Maschineller SprachÜbersetzung: Wissensrepräsentation und Textbezug", Niemeyer Verlag, Tübingen

Uchida, H. & K. Sugiyama (1980) "A machine translation system from Japanese into English based on conceptual structure ", in: COLING-80, Proceedings, Tokyo, S. 455-462