PRAGMATICS IN MACHINE TRANSLATION

Annely Rothkegel

Universität Saarbrücken
Sonderforschungsbereich 100
Elektronische Sprachforschung
D 6600 Saarbrücken
West-Germany

abstract

TEXAN is a system of transfer-oriented text analysis. Its linguistic concept is based on a communicative approach within the framework of speech act theory. In this view texts are considered to be the result of linguistic actions. It is assumed that they control the selection of translation equivalents. The transition of this concept of linguistic actions (text acts) to the model of computer analysis is performed by a context-free illocution grammar processing categories of actions and a propositional structure of states of affairs. The grammar which is related to a text lexicon provides the connection of these categories and the linguistic surface units of a single language.

1. The Problem

One of the main tasks of machine translation, besides the resolution of ambiguities and the generation of appropriate structural analyses, is the selection of adequate translation equivalents. It has been found that an analysis which even produces unequivocal results does not suffice for the production of pragmatically adequate texts in the target language.

There are problems with respect to the selection of appropriate lexemes, collocations, idiomatic expressions on the one hand. On the other hand we have to know what kind of syntactic patterns and anaphorical or elliptical constructions usually are applied with respect to the text type. What we need is information on cummunicative norms. In addition to a syntactic and/or semantic analysis we have to provide a pragmatic component especially in order to solve problems on the level of transfer.

The notion that linguistic usage and the selection of means of expression (lexis and syntax) is directed by - or at least influenced by - communicative intentions has received increasing attention with respect to problems of translation. Recent research in this area include communicative grammars for foreign-language learning (e.g. Leech/Svartvik 1975), but also more specific studies which explicity take account of text function (Reiß/Vermeer 1984, Thiel 1980) and aspects of action in texts (Hönig/Kußmaul 1982). These studies have influenced the theoretical foundations of TEXAN to the extent that we view communicative aspects as decisive for the solution of translation problems.

Some short examples of our texts (interacting-regulating texts, especially international treaties) may illustrate this approach. We should know when a special pattern has to be applied in different languages and when it has to be changed. It has been found in these texts that there is a special type of definition (DEFINE) with lexical restrictions and which always is realized by participle constructions in English, German, French, Italien, etc. A translation by a relative clause, e.g. in German, would be wrong. In a different text type it may be right or even must have this form. On the other hand, regulations (REGULATE) differ in verb forms. Thus in German present tense is to be used, in English shall-forms, and in French present and future may be alternatives. A general principle is, that the participants never are pronominalized.

The question now is what kind of linguistic model can help us to structure the relevant components of the analysis system?

2. Concept of Text Acts (TA)

Our system needs a linguistic model in which content, function and form of lingusitic expressions in a text are connected. We think that a good concept for this purpose may be the concept of text acts (Rothkegel 1984). TA are speech acts in which texts are produced. When we translate, we are producing a new text.

We follow Searle's analysis of speech acts into illocutionary, propositional and locutionary parts and assume, with respect to texts, the existence of three parts of text acts (I: text illocution; T: thematic specification of the propositional part; R: repertoires of lexical and grammatical expressions which are typically used for a specific communicative task).

$$TA : (I, T, R)$$

Automatic procedures for the processing of speech act basically have to do with the selection and representation of contextual factors. They determine the assignment of illocutions to linguistic utterances (Gazdar 1981). What models developed for this purpose have in common is the use of overall schemas whithin which the respective speech acts can be interpreted. While Evans (1981) handles general definitions of situation, Allen/Perrault (1980), Cohen (1978) and Grosz (1982) use general action plans in which the speech acts of interest are embedded. This principle, which is applied to dialogues in the models mentioned, we have applied to written texts in TEXAN (example of an article in Fig.1).

3. Model of Analysis

The analysis of text acts is oriented conceptually in a top-down fashion. In the context of machine processing, however, we have to rely on the linguistic surface as input data. TEXAN is a system which builds on other programs already completed within our project. We use a syntax parser (SATAN, cf. SALEM 1980), for instance, which provides a description of constituent structure and valencies. Furthermore, we use a program for case-grammatical analysis (PROLID, cf. Harbusch/Rothkegel 1984) which provides a role interpretation on the description of constituent structure. Input into TEXAN , then, is a complete structural and case-relational description of sentences. This determines

```
REGULATE (case)
├─ GENERALIZE (general case)
│   └─ FIX (activity)
│       ├─ INDIVIDUALIZE (partner)
│       └─ CONCRETIZE (commerce)
│           ├─ SPECIFY      (object,con)
│           ├─ DEFINE       (text)
│           └─ SPECIFY      (object,abs)
└─ DIFFERENTIATE (special case)
    ├─ FIX (condition)
    │   ├─ SPECIFY      (object,abs)
    │   └─ CONCRETIZE (event)
    │       ├─ LOCALIZE     (place)
    │       └─ INDIVIDUALIZE (partner)
    ├─ PERMIT
    ├─ FIX(activity)
    │   ├─ SPECIFY      (object,abs)
    │   └─ CONCRETIZE (commerce)
    └─ FIX ( condition)
        ├─ SPECIFY (object,abs)
        └─ CONCRETIZE (commerce)
            ├─ DETERMINE (relationship)
            └─ INDIVIDUALIZE (partner)
```

The Community	Die Gemeinschaft
shall not subject	führt (ein)
imports of products	für die Einfuhr der
defined under Article 1	in Artikel 1 genannten Erzeugnisse
to new quantitative restrictions.	keine neuen mengenmäßigen Beschränkungen.
If	
additional demand	Tritt (auf)
should arise	auf dem Gemeinschaftsmarkt
on the Community market,	eine zusätzliche Nachfrage,
the Community	so
will not object	hat ... nichts einzuwenden
	die Gemeinschaft, daß
to these quantitative limits	die vorgenannten Höchstmengen
being increased,	überschritten werden,
on the understanding that	sofern
the additional quantities	die zusätzlichen Mengen
shall be determined	von den Vertragsparteien
on the basis of mutual agreement	einvernehmlich
between the Parties.	festgesetzt werden.

Fig. 1

to a large extent the strategy of analysis within TEXAN. In priciple, the task here is to bundle the available information on syntax, lexis and thematic roles in a form suitable to the determination of the underlying illocution. Nevertheless, the concept of text acts is the basis for the structure of data. We distinguish the following components (Fig. 2):

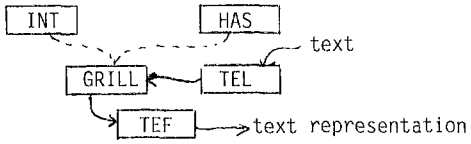

Fig. 2

The components of the automatic analysis are GRILL (grammar of illocutions), TEL (text lexicon) and TEF (sequence of propositions of the text). INT (schema of interpretation for the structure of states of affairs and communicative tasks) and HAS (action structure of the text) are preconditions in order to formulate the rules of GRILL. 'text' represents the input structure. This means that the sentences are syntactically analyzedand ordered according to a propositional listing. 'text representation' is output in the form of Fig. 1.

In the following we will sketch the structure of the components.

INT represents the structure in which knowledge of states of affairs is embedded into knowledge of linguistic action. It consists of 4 parts which can be combined. States of affairs (see Fig. 3):

(a) actions (a (x, (y), (z)))
states of affairs occur as actions/interactions
(a) of/between participants (x1, x2,...) and refer to an concrete object (y) or abstract object (z) or relate the two ones (y,z).

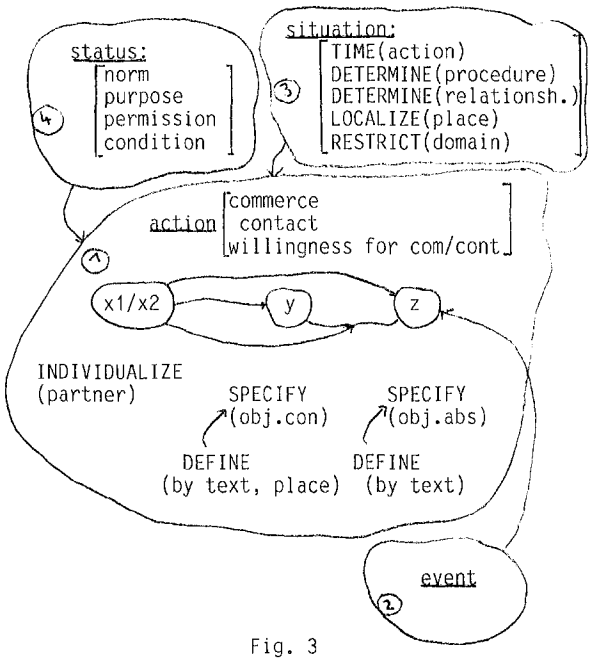

Fig. 3

(b) states of affairs occur as events concerning abstract objects: b (z)

(c) situation (m,n,o,p,...)
actions are embedded in a situation described by parameters of time, location, personal relationship, domain, procedures, etc.

(d) the verbalization of an action can be seen in the status of condition, norm, purpose, permission,etc.

Linguistic actions:
They are interpretations of states of affairs with respect to communicative tasks and can be described as predications on propositions. Thus we can add several types of illocutions to (a)-(d). Examples are:

```
CONCRETIZE (a (x, (y), (z)) )
FIX (condition (b (z)) )
```

HAS (Fig. 4) represents the action structure of 'treaties of trade' in terms of text acts. Our example in Fig.1 shows a segment of REGULATE (case).

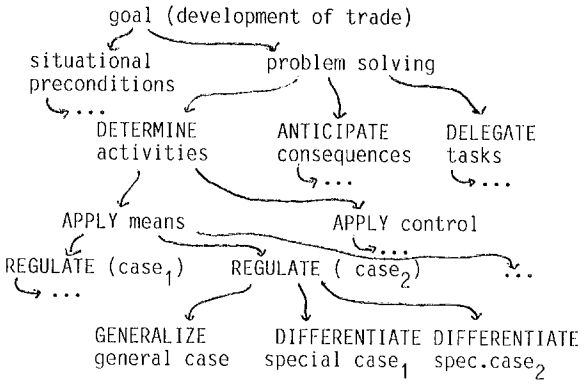

Fig. 4

TEL represents the text lexicon. According to the two tasks of TEXAN TEL includes two sections of information: an identification section concerning the text act structure (TAS) which is described by types of illocution and roles such as REGULATE (case), SPECIFY (object), etc., and a selection section consisting of lists of repertoires which belong to several single languages (TAE:R(L1,...,Ln). As a third part a key (K) is established which provides the connection of input data and the TA-information. On the level of simple illocutions the key represented by the lemma of the head of the respective phrase; on the level of complex illocutions the key is the illocution of a lower level. An entry of TEL has the following design:

```
TEL_i: 1. key (lemma or illocution_c)
       2. TAS (I/T)
       3. TAE (R (L1: l,g)
               R (L2: l,g)
                 ...
               R (Ln: l,g) )
```

It is possible that one key corresponds to several entries of TEL. This is the case if there are different TAS.

GRILL provides rules which represent the structure of INT and HAS and which transform them into procedures. GRILL (grammar of illocutions) has such a form that it can be processed by a context-free grammar parser. A parser has been developed according to the structure of the programming language COMSKEE. Elements of the **TEF**-component (listing of propositions of the text) are integrated as parameter (F) into the rules.

a) rule (R10) for terminals (lexicon rule):

$I_e\ (T_i)/(F')\ :=\ lemma_z,\ (T_i)\ /(F')$

e.g. CONCRETIZE (contact)/(F1) := "inform" (cont)/(F1)

b) rule for non-terminals (R1-R9), general form:

$I_c(T_j)/(F_{i-m})\ :=$

$I_l(T_f)/(F_i)\ +\ <\ |\ I_g(T_h),\ R\ |^n/\ (F_{o-p})\ >$

$||^n$ recursion
<> optional
R surface conditions

4. Transfer

On the basis of identified illocutions with respect to L1 we have access to the lexical and grammatical information of R with regard to L2, L3, etc. This information is offered by TEL. We apply a further assignment rule of the following type (e=english, d=german, l=lexical inf., g=syntactic inf.):

for 'lemma'(Lx), $I_i(T_j)$:= $R(l_f, g_k)$ (Ly)
for $I_c\ (T_j)$(Lx) := $R(l_f, g_k)$ (Ly)

Examples:

```
for 'subject'(e), CONCRETIZE(commerce) :=
                  R(l:'einführen', 'anwenden'
                    g: finite verb)(d)
for GENERALIZE (case) (e):=
                  R(g: main clause,activ,
                       present tense)(d)
```

The transfer part is to be seen as a kind of "helper" for translation purposes. It may be used by human translators as well as by systems generating the complete target text.

References

Allen, J.F./Perrault, C.R.,1980. Analyzing intention in utterances. Art. Intell. Vol 15,3,143-178.

Cohen, P.R.,1978. On knowing what to say: Planning speech acts. Ph.D.Thesis, Dep.of Computer Science, Univ.of Toronto.

Gazdar, G.,1981. Speech act assignment. In: Joshi,A. K./Webber, B.L./Sag, J.A., Elements of discourse understanding, 64-83, Cambridge, Univ.Press.

Grosz, B.J., 1982. Discourse Analysis.In: Kittredge, R./Lehrberger, J.(ed), Sublanguage, 138-174. Berlin, de Gruyter.

Harbusch, K./Rothkegel, A., 1984. PROLID . Ein Programm zur Rollenidentifikation. Ling. Arbeiten des SFB 100,N.F.8, Univ. Saarbrücken.

Hönig, H./Kußmaul, P., 1982. Strategie der Übersetzung. Tübingen, Narr.

Leech, G./Svartvik, J., 1975. A communicative grammar of English. London, Longman.

Reiß, K./Vermeer, H.J., 1984. Grundlegung einer allgemeinen Translationstheorie. Tübingen, Niemeyer.

Rothkegel, A., 1984. Sprachhandlungstypen in interaktionsregelnden Texten. In: Rosengren, I.(Hg.), Sprache und Pragmatik, Lunder Symposium 1984, 255-278. Stockholm, Almqvist & Wiksell Int.

--- 1985. Text Acts in Machine Translation. L.A.U.T. paper no. 133, Universität Trier.

SALEM. Sonderforschungsbereich 100 (Hg.), 1980. Ein Verfahren zur automatischen Lemmatisierung deutscher Texte. Tübingen,Niemeyer.

Thiel, G., 1980. Vergleichende Textanalyse als Basis für die Entwicklung einer Übersetzungsmethodik. In: Wilss, W.(Hg.), Semiotik und Übersetzen, 87-98. Tübingen, Narr.