[*International Conference on the State of Machine Translation in America, Asia and Europe*.
Proceedings of IAI-MT86, 20-22 August 1986, Bürgerhaus, Dudweiler]

Johann Haller, Dieter Maas

EUROTRA-D: Linguistic Research and Software Test Systems for Machine Translation

Table of Contents

1. Introduction

Basics of the Eurotra-D Work

The work in the Eurotra-D project is based on a variety of previous work and tries to evaluate and use the experiences gained in the research of fields related to MT.
"Traditional" linguistics (e.g. valency and dependency grammar as well as semantics) are as important here as modern grammar formalisms and syntactic theories (LFG, GPSG). It is worth noticing that traditional approaches deal more detailedly with a broader range of linguistic facts whereas grammar formalisms focus on a formal representation and aim at a generalisation of linguistic phenomena.

One of the principles of Eurotra-D is to use the experiences made by other former or running projects (METAL, SYSTRAN, LOGOS, etc), even if these are not designed explicitly on special linguistic models. They offer, however, a large amount of lexical and especially contrastive material and list problematic cases. Apart from this, studies on the application of these systems show which of their shortcomings and deficiencies are considered to be unpleasant from the point of view of the users.

The Eurotra-D project has access to the material and experiences of the SFB 100 (located in Saarbrücken) which has implemented various research oriented systems (SUSY) and which has compiled large dictionaries with e.g. more than 140.000 entries for German. These developments are partly being used (e.g. by MARIS).

2. Linguistic Research for MT

One of the principles in Eurotra is the stratificational approach. For analysis, this means that at three separate levels (or in three separate steps) information on configurational structure, dependency relations, and semantic relations is assigned. These levels are connected by "t-rules", i.e. translation rules. The levels which can be regarded as artificial representation languages serve as a starting point for linguistic research.

Each level is created by so-called "generators", i.e. grammars which should be developed independently from each other. The basic idea is the splitting of the very complicated translation relation between texts of different languages into several simple translation relations. The levels of representation between text1 and text2 should be motivated linguistically.

2.1. Syntactic Studies

In order to describe the syntactic level of Eurotra (ECS), studies are performed for the development of different syntactic analyses, some of which are based on the X-bar theories of Jackendoff and others.

An example for this are the rules for German main clauses by which all elements following the verb (or the subject in case of inversion) are subsumed under VPP:

```
cMCLl    = mcl, (npp, {case=nom}, v, vpp),
cMCLll   = mcl, (npp, {case=acc}, v,(npp,{case=nom}), vpp),
CMCL12   = mcl, (npp, {case=nom}, v,(npp,{case=nom}), vpp),
cMCL2    = mcl, (ppp, v, %advp, (npp, {case=nom}), v, vpp),
etc.
```

The underlying X-bar pattern is the following:

```
XPP   -->     (PRE-XP)  XP   (POST-XP)
XP    -—>     (PRE-X)   X    (POST-X)
X     -—>      lexical units
```

2.2. Valency studies

In order to describe the ERS-level, Eurotra-D uses as classification of valency frames which has been developed on the basis of the one used at the IdS in Mannheim. The following table gives the classification of the valency frames. The valency bound elements, i.e. the complements of the verbs are referred to as 'C'.

Table 1: List of Complement Classes

Class:	synt. realisation	sample sentence	verbs
C0 (NomCompl)	- NP, Nom, and PRO	(Fritz) schläft	nearly all not: Subj "es"
	- Daß-Satz	(Daß Du kommst,) gefällt mir.	
	- ind. questions	(Wer geht,) macht einen Fehler.	
	- infinitive-construction	(Gekommen zu sein,) war ein Fehler.	
C1 (Acc-Compl)	- NP, Acc	Er schlägt (ihn).	schlagen
	- Daß-Satz	Er sieht,(daß sie kommt).	sehen glauben
	- ind. question	Er sieht,(wen Du eingeladen hast)	kriegen küssen
	- inf. construction	Er glaubt (genug zu tun).	lieben hassen lesen
C2 (Gen.-Compl)	- NP, genitive	Er erinnert sich (des Fehlers)	sich erinnern
	- Daß-Satz	Er klagt ihn an, (daß er einen Mord begangen hat)	gedenken anklagen überführen
	- ind. question	Er beschuldigt ihn, (wessen er sich erinnert).	bedürfen sich enthalten
	- inf. construction	Er beschuldigt ihn, (einen Mord begangen zu haben).	
C3 (Dat.-Compl)	- NP, dative	Ich helfe (ihm).	trauen helfen

	- ind. question	Ich traue, (wem ich will)	gefallen geben nützen
C4 (Prep.-Compl.)	- PP	Ich warte (auf ihn).	warten auf sich er-
	- ind. question	Ich warte,(worauf mein Freund auch wartet).	innern an
	- Daß-Satz w. corr.	Ich warte(darauf, daß er kommt).	wählen unter, zu
	- Inf.-constr.	Ich warte (darauf,eine Chance zu haben).	sich wa- gen an verstehen von

There are the following additional complements:

C5 (locative complements)
C6 (temporal complements)
C7 (directional complements)
C8 (predicative complements - nominative, accusative, as adjectival groups, etc.
C9 (infinitival construction)

These valency classes which are used in a similar way for nouns and adjectives are intended to filter out those readings which are not realized syntactically when translated from ECS to ERS.

Verbs are classified according to 4 different process types: relational, mental process, communication, and action. The most important possible arguments of the process types are: Agent, Attribuant, Identifier/Classifier, Location, Processor, Phenomenon, Sender, Message, Receiver, Range.

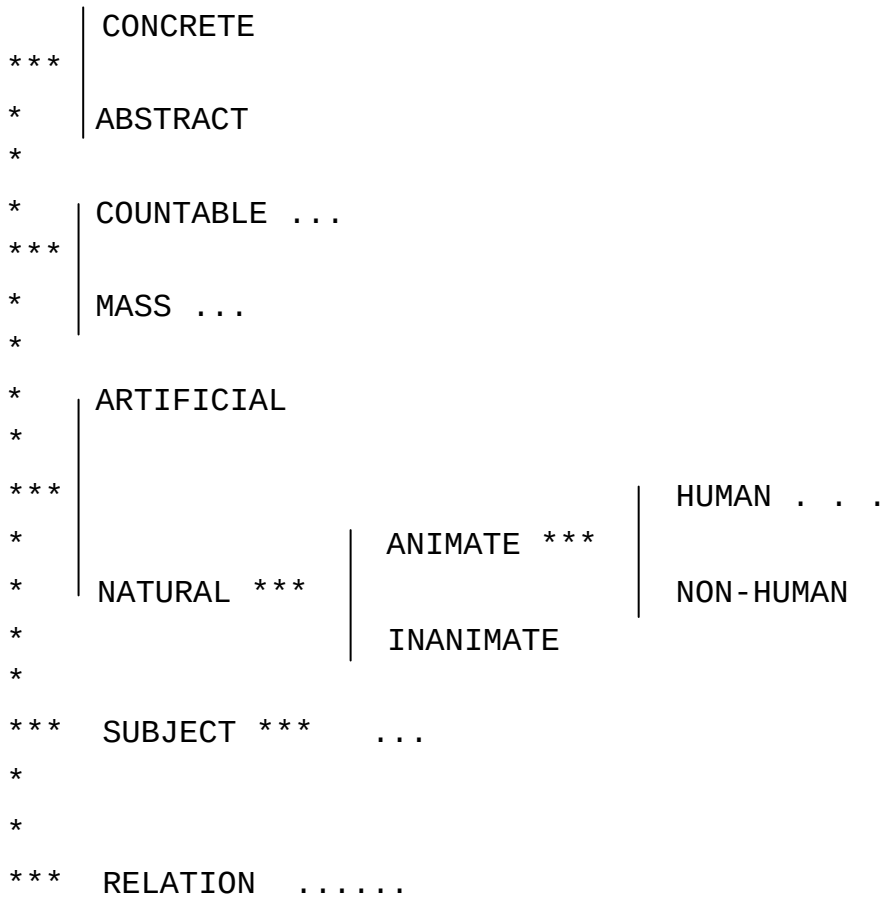
Theoretically, it would be possible to differentiate readings on the basis of complement classes and their semantic categories only; however, introducing semantic relations simplifies the transfer module in a multilingual system essentially.

2.3. Semantics

For a discrimination of the argument names (labels) of the semantic relations as well as for an even more precise description of the possible translations, an operationalizable system of semantic categories is required. The number of these categories has to be limited and they have to be ordered hierarchically. The categories have to be motivated by an extra-lingual perception such that the basic classification for all (Euro-) languages remains unchanged. A language dependent Refinement in individual branches of the category tree is possible.

In the current state of our research system for semantic categories of entities, the most important category pairs to be assigned alternatively in this sequence are the following:

Table 2: Principal Semantic Categories for Entities



In addition to these, features for special subject fields which are generally required for disambiguation are assigned under SUBJECT.
Operationalization is performed by a coding program which takes into account the internal dependencies of the category tree.

3. Development and Application of Software

3.1. Eurotra Software: The <C,A>t-formalism

Different implementations of the Eurotra formalism run on a SUN-III under UNIX. At the moment we have a morphological analysis of German covering nearly all morphological phenomena, and an ECS-grammar, i.e. a grammar of the constituent structure, covering a restricted range of phenomena as defined in the intensional Eurotra corpus definition, as well as some first outlines for the ERS- and IS-structures.

At the moment we are adapting our grammars in the Eurotra formalism to version 1.2 of the Eurotra software. The most important extension is the introduction of two different types of rules: b-rules, i.e. the building rules which build the structures, and a-rules which test the correctness of the structures built (e.g. agreement). The advantages and disadvantages of this version will be discussed below. Some of the advantages are the improved user-friendliness, and better debugging and testing facilities. Improved running times cannot be expected here due to the underlying philosophy of the purely declarative system and its implementation in C-Prolog.

3.2. A Multilingual Dictionary System

For the compilation of the transfer dictionaries we have a dBase-III development (University of Bonn, IKP) at our disposal which is designed to facilitate the coordinated construction of a Multilingual dictionary.

This tool can perform the following tasks:

- compilation of bilingual transfer dictionaries
- intersection with text retrieval systems to help in disambiguation
- interlingual consistency between various transfer dictionaries
- quick usage of transfer dictionaries in prototype software
- easy exchange of dictionaries between different language centers

3.3. SPES

3.3.1. Introduction

The author of SPES, member of the former "Eurotra-Software Group" and being in charge of several Eurotra contracts on software and user language, decided in early 1981 to implement a software system on the basis of the theoretical Eurotra developments aiming at a processing system which would be able to give evidence on its behaviour as well as on the user's needs. The work on this prototype resulted, however, in a whole series of prototypes, each one being an improvement of its predecessor. As all these prototypes were connected to certain modules of the translation system SUSY – its dictionaries and its morphological analysis – a rather stable state within the prototyping development was called SUSY-II. An improved version of SUSY-II has remained one of the most important components of SPES.

When the EUROTRA Central Team in early 1985 abandoned the former Eurotra framework by introducing the <C,A>t philosophy – on which the current Eurotra software is based now – the author of SUSY-II decided to enlarge the power of the system by introducing a translational component based on Eurotra's ideas of t-rule and compositionality. Only now the system, which was called "Saarbrücken Prototype of the Eurotra Software", became capable of translating – not only between different languages, but also between different levels of representation.

3.3.2. Theoretical Background

Description of a 'level'

Although SPES has been developed in close contact with the development of the Eurotra theory, it is not at all simply another implementation of the same ideas.

Just like in Eurotra, we can define different levels of representation in SPES, and it is up to the linguist, which levels he chooses and which concrete linguistic legislation he prefers, as long as he remains in the boundaries of the software system.

In SPES each level is described by a process system which consists of processes arranged in the form of a tree. The processes contain, as preterminal elements, so-called grammars which themselves are collections of rules. Processes and grammars may have entry and exit conditions, parameters for creation, stratification, preference, etc. By means of the process system, the linguist can control the application of the rules to a large extent.

The rules themselves have the shape of context free rules, consisting of a left hand side (lhs) and a right hand side (rhs). The lhs describes a sequence of 1 to 4 objects, which describe partial trees (always rooted) or single roots. The rhs describes what the resulting object will look like, when the lhs matches a piece of actual data structure. The formalism can be used for the description of CF rules, by simply not using its full power.

lhs and rhs describe only the geometrical shape of the trees. As each node of a tree is associated with a set of features (property-value pairs) the rules contain "conditions" and "assignments". The conditions are boolean expressions built on terms which express relations between features of the nodes mentioned in the lhs. If the evaluation of the conditions yield the value "true", the rules fires, i.e. it produces a new tree. The assignments transfer features from the nodes, specified by

the lhs, to the new structure, normally to its root.

Processing starts with the top node of a process by traversing the process tree downwards and from left to right. When the top node is reached again, the process stops. At this point the results are checked against the process goal, which normally results in discarding unwanted structures.

The filtering device has a built-in preference mechanism: If there is no structure which meets the goal requirements (and the 'goal' results are always single trees, not sequences), the system selects those sequences which seem most promising. Therefore, even if analysis fails, the system returns some result(s), normally a sequence of trees, which are not connected to a top node.

Description of a translation

When the processing system has managed to produce a description of an input sentence for a level of representation L_1, we can translate this into a description on level L_{i+1}. Each node of an L_i-tree contains the information by which rule it was built (all rules have unique names!). Therefore, the translation rules consist of a lhs and a rhs, the lhs being a name of a rule of L_i. The rhs consists either of a sequence of names of rules of L_{i+1} (together with application parameters, like "try them all", "stop after firing") or a name of a process of L_{i+1}. There may exist additional rules which do not build structures, but operate only on features.

The semantics of such a translation rule is the following:

If node n was built by application of lhs to its daughters $d_1,...,d_r$, then the target tree(s) will be built by applying rhs to the translations of $d_1,...,d_n$. Because of this procedure, a tree of the source level L_i may have several translations in L_{i+1}. Therefore we can say that the translation produces local ambiguities which however then should be filtered out by the process system of the target level L_{i+1}.

The translation works either bottom up or top down, but bottom up is the normal case. In the top down mode the user can specify for each new daughter node a rule name, which is then looked up in the set of t-rules providing the rules or processes which have to expand it. By doing this, the software system uses the tree under construction just like a process system.

3.4. SPES and the EUROTRA-Software

3.4.1. The Eurotra Prototype Version 1.2.

The actual prototype operates in the following way:

- A sequence of sentences is read in from a file.

- The user selects one (or several) sentence(s), and this sentence is segmented into wordforms (blank as separator).

- The single words are looked up in the "lexicon". The result is stored in a chart.

- The first level of representation is built according to a rule system by using the Early algorithm.
  There are two types of rules:
  a) b-rules: context free rules, which describe the structures

b) a-rules: they check or percolate features, using the unification mechanism.

The system will find all interpretations according to the rules for an input sentence. If the s-symbol cannot be reached, there will be no output.

- Translation to the next level by using t-rules for structure building and feature transfer.

A t-rule consists of an lhs and an rhs. The lhs is a description of a piece of structure, a tree of arbitrary depth, in which nodes can be identified by arbitrary local names. The rhs describes a tree as well, but its nonterminal nodes being names of b-rules of the target level. The local names of the lhs reappear, if at all, as terminals of the rhs tree.

There may be a whole sequence of such translation steps.

3.4.2. Differences between SPES and EUROTRA

SPES does not use unification: checking agreement e.g. does not have any consequence, except the user states the consequences it should have.

Eurotra has no control system, its rules are not grouped, and the user has no possibility to guide their application: extreme non-determinism. SPES enables the user to control the rule application, specify preferences, i.e. having determinism to some extent, especially for building the first structural level.

Translation rules in Eurotra are totally deterministic: Each rule produces one or no result. The target level is degenerated to a simple enumeration of rules (which of course must fit together). The structure building is therefore rather deterministic. The

main task of the target level is checking, not building, because in fact the t-rules describe the resulting object rather precisely. In SPES however, a whole rule system (a process) can **be** applied to translations of daughters for producing the translation(s) of a tree: The target level operates non-deterministically as well, if wanted.

The user languages of both systems are very different. SPES' processes, grammars, and rules are formulated in a language which is quite near to former Eurotra proposals (cf. Maas & Maegaard: Syntax and Semantics of the Eurotra Formalism). The actual Eurotra formalism is considerably more concise than the SPES language. But on the other hand, the SPES formalism is considerably more powerful.

In SPES, rules can be compiled individually, e.g. after corrections, and also single grammars can be compiled individually. Only in extreme cases a whole process system must b**e** recompiled. This makes debugging and enlarging the system rather comfortable. In the Eurotra prototype, a level has to be reinstalled as a whole after changes, which takes a lot of time.

SPES has an interface to dictionaries (of the SUSY type), which allows to describe a dictionary access within a rule (in the conditions part). In the Eurotra prototype, there is only a lexical component, which assigns features to wordforms, before structure building starts. The lexicon cannot be accessed during structure building. However, a dictionary can be simulated through b-rules or a-rules.

3.4.3. The Actual State of SPES

In this section we will sketch how the SPES system actually runs.

- The input text is processed by the SUSY module LESEN. The result is a sequence of word forms (with typographic information etc.).

- The SUSY module WOBUSU treats morphology (inflection, composition, fixed phrases).
  Only after this SPES starts to operate.

- Production of a constituent structure.
  After this the translation processes start.

- Production of a relational structure (relations and case frames according to P. Schmidt).

- Refinement of the relational structure by using semantic features. Relations unchanged.

- Production of relational structure of the target language by translating lexical units together with their case frames.

- Refinement of relational structure by using semantic features of the target language.

- Production of a configurational structure.

- Production of final constituent structure together with inflection. Projection of the terminals (strings) shows the translation of the sentence(s).

For the production of the relational structures special dictionaries (resp. dictionary entries) are needed, which describe the case frames of the lexical units (especially verbs). Therefore only some examples can be translated up to now, although the grammars for the different levels are quite extensive. Only the production of constituent structures (for analysis) will produce results for any input.

4. Concluding Remarks

As has been lined out here, it is the task of EUROTRA-D, the German unit of the European research project EUROTRA, to perform linguistic research on theories and material which are to be applied in MT. The procedures worked out on this basis are then implemented for a testing in different software systems.

During this phase, information is exchanged, and discussions and experiments on linguistic research concerning analysis, transfer, fend generation as well as on software problems take place with the members of Eurotra-D's parallel research units (Universities of Stuttgart, Berlin, Bielefeld).

The results worked out in the national group are then introduced into the EUROTRA project via the different cooperation institutions. International cooperation in EUROTRA is performed for example by current joint contrastive studies on the IS-structure with the Danish, English, and French language groups; D. Maas is a member of the group working on software concepts and J. Haller is a member of the Liaison Group. A larger project on a multilingual description of predicates is planned which will be coordinated by E. Steiner.

In spite of the many (Euro-) bureaucratic problems, some essential results were achieved which will be shown in the planned demonstration.

5. References

For further information and references, cf. the respective chapters in:

- The EUROTRA Reference Manual, Version 1, Revision 2

- EUROTRA-D, Erster Zwischenbericht zum Jahresende 1985, Förderungs-Nr: 1013208 1

- Heinz-Dieter Maas, Das SPES-Handbuch (forthcoming)