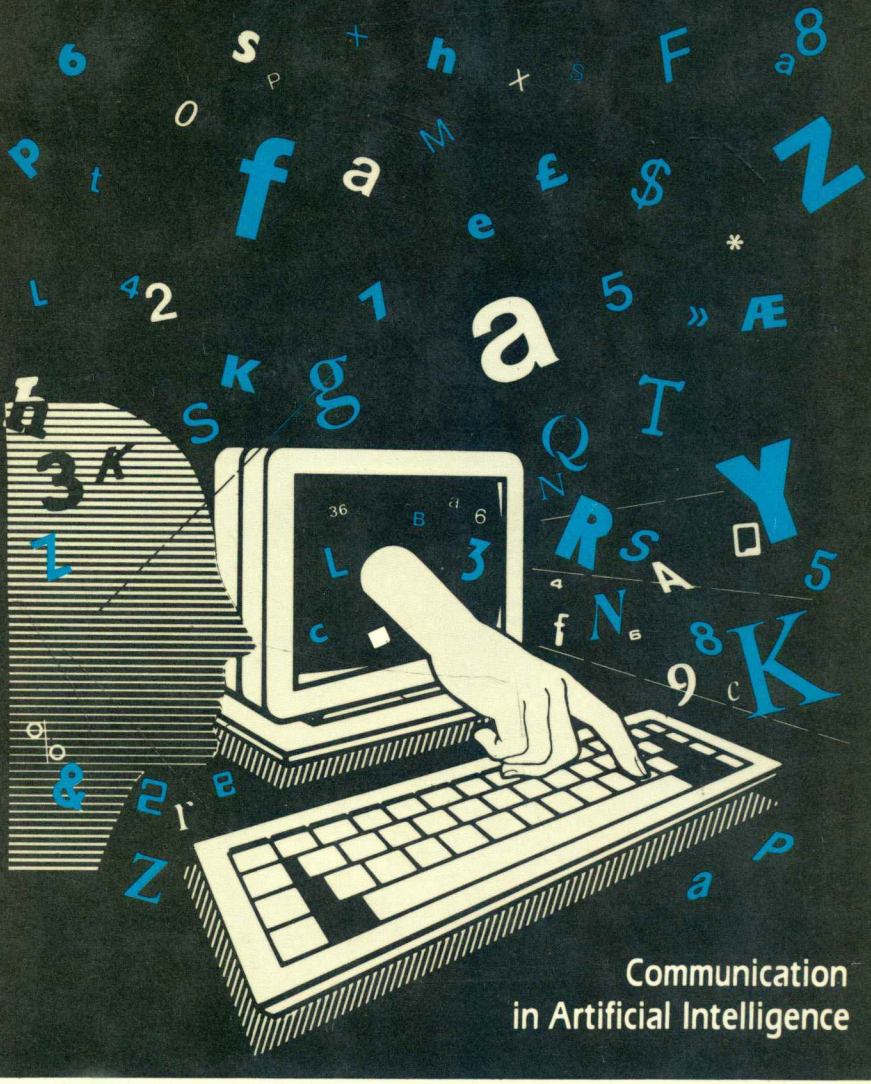
FROM SYNTAX TO SEMANTICS

Insights from Machine Translation



Communication in Artificial Intelligence

EDITED BY

Erich Steiner, Paul Schmidt and Cornelia Zelinsky-Wibbelt

From Syntax to Semantics

Insights from Machine Translation

Edited by
E. H. Steiner, P. Schmidt and C. Zelinsky-Wibbelt

江苏工业学院图书馆
藏书章

Pinter Publishers, London

© E. H. Steiner, P. Schmidt and C. Zelinsky-Wibbelt, 1988

First published in Great Britain in 1988 by
Pinter Publishers Limited
25 Floral Street, London WC2E 9DS

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted by any other means without the prior written permission of the copyright holder. Please direct all enquiries to the publishers.

British Library Cataloguing in Publication Data
A CIP catalogue record for this book is available from the British Library.
ISBN 0 86187 960 0

Typeset by Joshua Associates Ltd., Oxford
Printed by Biddles of Guildford, Surrey

Contents

试读结束，需要全本PDF请购买 www.ertongbook.com

1 Introduction

JOHANN HALLER, PAUL SCHMIDT, ERICH STEINER, ELKE TEICH AND CORNELIA ZELINSKY-WIBBELT

It is our hope that the contributions in this volume will be of interest to readers who are computational linguists, linguists, and translators—of interest from two points of view.

First, the core of the book describes an implemented system for the analysis and synthesis of German in a multi-lingual machine translation (MT) environment (Chapters 2, 3, 4, 7, 9 and 11). The remaining contributions explore theoretical and practical questions arising from work on this system, yet leading on to possible alternatives. The theoretical heterogeneity, which readers will find in this book, is largely due to the practical and inescapable demands of an MT system, which ultimately is designed to translate between all of the official EC languages, a system, at which currently well over a hundred people are working. This presupposes the ability to reach agreement on a theoretical and on a practical level, an agreement which, as in our case, has to lead to a running system.

Second, a considerable part of the work described here, though not all of it, is part of the work of the German language group of the EUROTRA project, which will be described below in this introduction. It is important to be aware of this organizational environment for the system described here, because this was the main reason why the authors represented in this volume had, and still have, the opportunity of working with each other, the only exception being the EUROTRA-D/SEMSYN co-operation, which is not part of the EUROTRA project as such. In a very real sense, then, all of the contributions in this book have a close interrelationship. So, while this is not some official report on EUROTRA work, almost all of the ideas presented here have had an impact on the project.

As we said above, the authors of this volume are used to working in a multilingual environment. Therefore, great care has been taken to ensure that, although essentially we are referring to a system for the analysis and synthesis of German, everything we say is exemplified on, at least, English material, but frequently also using examples from other European languages.

Before proceeding to the individual contributions in this volume, we would like to give an outline of the EUROTRA project, which will then be followed by a brief abstract-like survey of all the contributions in the book.

EUROTRA is a joint project of all twelve member states of the European Community. In the first place, EUROTRA is designed to promote both computational linguistics in general and knowledge transfer in the EC in particular. In addition to that, a preindustrial prototype is planned which by 1990 is supposed to translate texts written in the administration language of the EC. All nine languages of the Community are to be covered both in analysis and generation, and transfer in each of seventy-two translation directions has to be worked out.

The Commission of the EC is the organizer of this project. They are financing the central research in linguistics and software and are contributing to the cost of the national components. Thus, the goal of the EUROTRA project is manifold: on the one hand, it is to stimulate research in the area of computational linguistics and the inner-European exchange and transfer of knowledge; on the other hand, it is expected that a prototype of an MT-system which will actually be used will be developed.

During the preparatory phase (1978–84) progress was only made towards the first goal; since the beginning of 1985 the second goal of the program of work has also been explicitly defined—this goal, however, has not come very much closer for organizational reasons (several countries joined the project only recently and the Commission's team has not yet been provided with sufficiently competent staff). At the end of the current—second—phase a prototype with '2,500 dictionary entries' should be completed, subject to a critical evaluation. Along with the evaluation of the politically and scientifically achieved goals this forms the basis of decision for introducing the third phase and for rating the chances of realization in general—if there is no redefinition towards a pure research and education project in the meantime.

1.1 ORGANIZATION: EUROTRA-D AND SUBSIDIARY RESEARCH

Along with the Saarland, the Bundesministerium für Forschung und Technologie (BMFT) contributes 52 per cent to the financing of the German component EUROTRA-D at the Institut zur Förderung der Angewandten Informationswissenschaft (Grant No. 1013208/1). The task of the German research group in Saarbrücken is the development of the German components of analysis and synthesis and the dictionaries; the transfer dictionaries (from the other official languages to German) are being developed in co-operation with the University of Bonn.

The grammar formalism that is used by all the language groups has been developed by a central team. This formalism follows the principles of unification grammars. Another basic principle is that the translation process is split up into several smaller parts. The basic assumption here is that between source text and target text a number of intermediate representations may exist between which translations take place. For German, descriptions on morphological, configurational and relational levels are being worked out

based on current research results in syntax and semantics. At the same time, German researchers are working on the development of the conception of the EUROTRA-formalism and are participating in the work on problems connected with efficient implementation.

To support EUROTRA-D the BMFT sponsors independent subsidiary projects at the universities of Berlin, Bielefeld and Stuttgart. The particular tasks of these projects consist of testing and rendering useful the results and methods of theoretical linguistics for application in machine translation. Hauenschild's contribution to the volume deals with GPSG and its possible application in this context; the project in Stuttgart is represented by Chapter 8 by Eckert and Heid.

1.2 THE CONCEPT OF LINGUISTIC LEVELS

The design of the EUROTRA-project for the development of an MT-system implies that the translation from source language to target language is split up into several linguistically motivated levels of representation. The level of semantic representation at which transfer from one language to another will take place is the 'Interface Structure' which is to show 'euroversal' design which is NOT an interlingua. The EUROTRA Reference Manual (Arnold, des Tombe & Jaspaert 1985) contains the official legislation for the representation of each level and is obligatory for each language group.

1.3 RULE FORMALISM AND SYNTAX

The representation language in EUROTRA is the so-called C,A,T-formalism (Arnold, des Tombe & Jaspaert 1985). It works with three objects:

Constructor
Atom
T-rule

and it is based on a stratificational theory of the translation process, which is:

(1) the translation relation is not defined directly between source text and target text but has to be split up into several simpler steps;
(2) these steps are to be carried out between linguistically motivated levels.

From these assumptions three main features of the EUROTRA formalism follow:

(1) The translation system has a stratificational design: the translation relation is between T1 ... R1 ... Rn ... T2, where each level R is an artificial representation language. At the moment Ri = {EMS, ECS, ERS, IS} holds.

    EMS = EUROTRA Morphological Structure
    ECS = EUROTRA Configurational Structure

ERS = EUROTRA Relational Structure
IS = Interface Structure

(2) The representations are determined by grammars, so-called 'generators'. These contain two types of rule:

b-rules building representations;
a-rules expressing generalizations over attributes.

(3) The relationship between the two levels is determined by a 'translator' consisting of a number of 't-rules'. These t-rules have the following characteristics:

— they are one shot, i.e. they do not have an internal strategy;
— they are compositional, i.e. the translation of a structured object is a function of the translation of its parts.

Thus, on the ECS-level rules have to be formulated for a syntax analysis of German which reflects the state of the art in research. The rules are based on the relevant literature in the field (Reis 1985; den Besten 1983; etc.): from a canonical word order (finite verb in final position), all variants are derived by 'movement-rules'.

The relational structure, or dependency structure, which corresponds to the f-structures in unification grammars, is defined by the property of the lexical units to subcategorize for other elements. This property is called 'valency'. The definition of the relational level for German is based on works of the IdS (Institut für deutsche Sprache), Mannheim. The German ECS and ERS are presented in Schmidt's contribution to this volume (Chapter 2).

1.4 THE EUROTRA INTERFACE STRUCTURE

The Interface Structure, from which transfer to the target Interface Structure is to be carried out, is described in Chapter 4.5 of the EUROTRA Reference Manual with the heading 'Interface Structure and Transfer'. The Interface Structure (IS) in EUROTRA is defined as a level of minimal transfer between source language and target language. Thus, IS is not an interlingua, which is a fundamental characteristic very often overlooked in the discussion about EUROTRA. The theory of IS comprises a number of component theories: MODALITY, TIME, SEMANTIC FEATURES, SEMANTIC RELATIONS and maybe others.

Semantic Relations (SRs) are the semantic relationships between the 'governors' (govs) of a construction and the members dependent on them, the complements (comps). Semantically speaking, these are the relationships between predicates and arguments (preds and args).

One of the contributions to this volume, Chapter 3 by Steiner, Eckert, Roth and Winter-Thielen, provides a survey of the proposals of EUROTRA-D for the inclusion of this component theory; another contrasts these suggestions to

the modern LFG-formalism, and the contributions of Heid, Rösner and Roth (Chapter 6) and of Eckert and Heid (Chapter 8) discuss the corresponding applications both in generation (with SEMSYN) and in the production of transfer lexicons; Zelinsky-Wibbelt's contribution (Chapter 4) deals with the semantic feature system. On the level of the Interface Structure sentences are supposed to be represented as 'euroversally' as possible, meaning that the transfer steps between the Interface Structures of the various languages shall be kept as small as possible. Here, the different readings of a lexical unit as well as of structural constituents are identified by means of semantic features. The structural starting points are so-called 'deep syntactic relations'; ideally, transfer is reduced to the transfer of a lexical unit from source language to target language (which is, however, very often impossible to achieve).

In an attempt to substantiate the term TRANSFER in model-theoretic terms, a translation is then called 'acceptable' or a 'q-paraphrase', if the pair

t_0, and r_0 and t', r

is true in the same possible worlds of source and target language, where t stands for the sentence and r stands for a particular interpretation of a sentence.

For the structural definition of IS the distinction of (bound) ARGUMENTS and (free) MODIFIERS is relevant; a word is to have several readings if different argument structures can be assigned. Figure 1.1 shows an example of an IS-representation of a simple sentence.

Die Industrie wird in Europa seit 1980 verbessert.
(The industry has developed in Europe since 1980.)

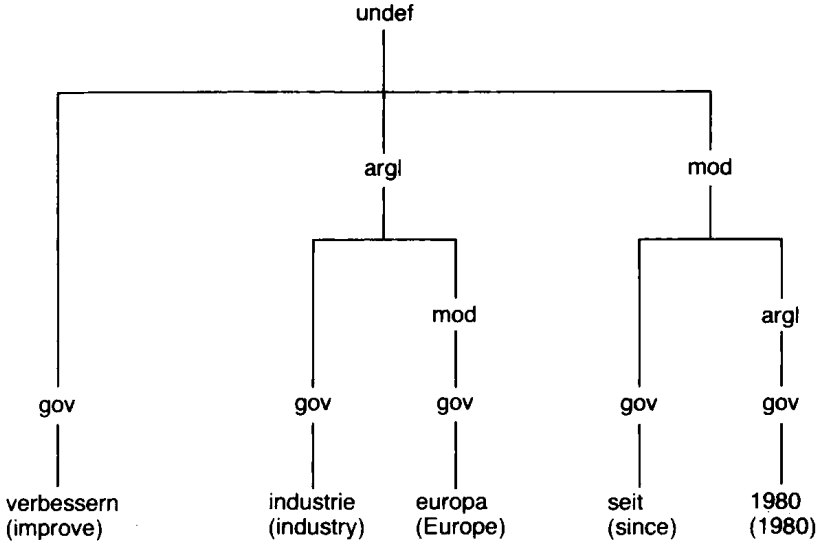


Figure 1.1

试读结束，需要全本PDF请购买 www.ertongbook.com

TERMINAL NODES are word forms on surface level with the following exceptions and peculiarities:

- only forms without inflectional endings (some IS-features have been derived from the morphology, e.g. tense, number);
- auxiliaries have been changed into features;
- articles are expressed as values of the feature 'defs' (definiteness);
- multiple word expressions and idioms have been put together in one unit;
- separable prefixes have been concatenated with the verb;
- prepositions with valency-bound PRPs have been stored as values of the feature 'pform';
- conjunctions introducing an argument clause have been stored as values of a corresponding feature 'subconjform';
- empty subjects like the English *it* (is clear that) and German *es* have been deleted, the same holds for infinitive particles (G, *zu*, D *at*).

As already mentioned, the IS will contain quite a lot of features, which are used both for monolingual disambiguation and also handed over in transfer as a means of providing for correct generation.

1.5 PROBLEMS AND PROPOSALS

In the development of every machine translation system the question about a systematic procedure for the development of motivated sets of feature-value pairs arises at some point. After several linguistic investigations had been carried out in the preparatory phase of the EUROTRA project (which did not result in a list everybody agreed with), it was decided to start with a relatively scanty description of the IS: the IS is basically linguistically motivated and semantic information will be incorporated in the above-mentioned cyclic procedure. Some first suggestions with respect to the following issues are made in the Reference Manual:

- argument structure: for the time being the arguments are numbered (ARG1–4) and have no semantic meaning whatsoever (except for order and for the variation of the English indirect object): 'give the baby a toy'—'give a toy to the baby';
- problems with a few areas are discussed (reflexive constructions, arguments with nouns and adjectives);
- loose connection of modifiers;
- three feature-value-pairs for semantic categories (HUMANNESS, ABSTRACTNESS, COUNTABILITY);
- definition of research tasks for TIME, SPACE, DEFINITENESS (placing of articles); DIATHESIS (active/passive).

The further procedure of developing the Interface Structure is supposed to be 'cyclic', which means that in the first place evidence from the corpus should be

collected and then a translation should be provided either by comparing the word or expression with the corresponding passage in the corpus version of the other language or otherwise not using the corpus. Whenever the occurrence in the respective subject area seems probable a translation shall be specified (including different readings of ambiguous words and structures). This method is exactly the one that has so far been used in developing machine translation systems; the consequence in systems like SYSTRAN, METAL and LOGOS is that at different points of time information for the same units is produced and incorporated into the linguistic knowledge base. Before long problems will arise that are identical with those of the existing systems.

Compared with that, EUROTRA-D recommends making use of this experience and starting with an ordered set of basic features right from the beginning; in order to do so, an experimental phase with several languages and transfer experiments is necessary. This experimental phase can be carried out more systematically if a relatively complete theory of the description of semantic phenomena can be made use of.

The models described in the first three chapters (Schmidt, Zelinsky-Wibbelt, Steiner *et al.*) are implemented in the form of a runable system (in the EUROTRA-formalism), which was demonstrated in the spring of 1987 when the system was evaluated by an advisory board of linguists. There are also experimental implementations with the other component theories, which are to be integrated in the near future. All approaches introduced in this volume are to be considered as proposals by one national group within EUROTRA, and they may have a long way to go before actually being applied in several or all language groups. It is for this reason that all implementations cover analysis/generation up to Interface Structure only; at the moment this cannot be extended to other languages. However, as this book goes to press, organizational preparations towards this step are under way.

Before bringing this introductory section to a close, let us give a brief overview of the contents of the individual contributions in this volume.

Chapter 2: A syntactic description of a fragment of German in the EUROTRA framework (*Schmidt*)

This chapter takes up certain issues of the syntax being used in the project, which is similar to a functional framework like LFG, and the interaction between the syntax and the formalism. It also illustrates how standard problems of German syntax can be treated within the given framework.

Chapter 3: The development of the EUROTRA-D system of Semantic Relations (*Steiner, Eckert, Roth and Winter-Thielen)*

Taking the representations discussed in the preceding chapter as an input, this chapter discusses some of the major linguistic requirements for a system of Semantic Relations within the project. It argues for the particular suitability of ideas from Systemic Functional Grammar (SFG) for this task, relying heavily

on work by Robin P. Fawcett published in SFG. At the same time, an analysis of a large number of German clause patterns in terms of Semantic Relations is presented.

Chapter 4: From cognitive grammar to the generation of semantic interpretation in machine translation (*Zelinsky-Wibbelt*)

Semantic features are, apart from the SRs discussed in the preceding chapter, an essential component of the semantic representations being used in the overall analysis. We give a set of features, explain their theoretical basis and operationalization, and illustrate the implementation of these features. The theory of semantic features is formalized as far as necessary in the context of the present volume.

Chapter 5: Semantic Relations in EUROTRA-D and LFG: a comparison (*Steiner*)

Here we take up again the system discussed by Steiner *et al.*, contrasting it with the approach by LFG in the area of Semantic Relations, or Thematic Roles. This chapter highlights some of the formal characteristics of the present system. It also discusses a range of standard questions in its area, thus being of interest not only in our specific context, but also for a comparison of LFG and SFG in general. The scope for theoretical depth is, necessarily, restricted by the overall purpose of the present volume.

Chapter 6: Generating German from Semantic Relations: Semantic Relations as an input to the SEMSYN generator (*Heid, Rösner and Roth*)

This chapter reports on the implementation of the Semantic Relations component illustrated by Steiner *et al.* into the SEMSYN system for text generation. While there is an intimate connection between this chapter and the chapters by Steiner *et al.* and Eckert and Heid, it should be of independent interest for text generation.

Chapter 7: Transfer strategies in EUROTRA (*Schmidt*)

This chapter presents some strategically motivated investigations into the problems of transfer in a multilingual translation system as depicted in this volume. It makes extensive use of the ideas presented in the chapters by Steiner *et al.* and Zelinsky in this volume.

Chapter 8: Semantic Relations in EUROTRA-D and syntactic functions in LFG: a comparison in the context of lexical transfer in machine translation (*Eckert and Heid*)

This chapter, again, takes up the SR system developed in earlier chapters investigating the function of Semantic Relations in lexical transfer from source

language to target language. Specifically, the present approach is contrasted experimentally with an approach using only LFG syntactic functions in transfer.

Chapter 9: The transfer of quantifiers in a multilingual machine translation system (*Zelinsky-Wibbelt*)

In this chapter we make a proposal for a 'euroversal' semantic representation of quantification, which is developed under special consideration of transfer between different expressions designating analogous set properties of an entity. As this chapter deals with the quantification of nouns, the proposal is intricately related to the semantics of nouns which is developed in Chapter 4.

Chapter 10: A constructive version of GPSG for machine translation (*Hauenschild and Busemann*)

The question discussed here is that of the applicability of Generalized Phrase Structure Grammar (GPSG) for MT. There are two major subsidiary tasks:

- Defining the role of GPSG within a full-fledged MT system;
- Creating a constructive version of GPSG theory from the classical version (Gazdar *et al.* 1985), which is purely declarative, as a prerequisite for an efficient implementation.

This chapter, as well as those discussing LFG relative to the proposals made in the rest of the present volume, aims at investigating some major contemporary linguistic theories against the background of a multilingual MT system. This chapter interacts strongly with Schmidt's chapter on syntax, thus closing the circle from syntax to semantics, transfer and generation back to syntax.

Chapter 11: LFG and the CAT-formalism (*Schmidt*)

In this chapter a proposal is made for an improvement of the formal basis of the translation system depicted here which would solve some problems arising in the area of German syntax. The ideas proposed are 'imported' from a certain version of LFG.

After giving the necessary theoretical and organizational context of this volume, we would now like to invite the readers to enter the discussions themselves and to explore the system which we are describing. The authors hope that in writing the contributions to this book, they have managed to avoid both the extremes of being over-technical and being too superficial and undemanding. The editors would finally like to take this opportunity to express their indebtedness to all those who have made this volume possible, whether as authors or in other helpful ways.

Part I A coherent system—theory and implementation

2 A syntactic description of a fragment of German in the EUROTRA framework

PAUL SCHMIDT

2.0 INTRODUCTION

This chapter describes the syntactic part of the German module of the multilingual translation system sketched in this volume. The grammar in the CAT format (CAT = Constructor Atom Translation rule) given here has been implemented as the German part of the EUROTRA project. The following description of this implementation divides into two parts:

(1) In a first part, there will be some introduction of the theoretical basis underlying the descriptions, basics of an empirically linguistic kind and basics of a formal kind, concerning the formal power of the representational language developed for the representation of the linguistic facts.
(2) In a second part, it will be shown how a fragment of German has been described on the basis of the theoretical givens depicted in section 2.1.

2.1 THE REPRESENTATIONAL LANGUAGE

2.1.1 Basics of a theory of automatic translation

According to the principles adopted for EUROTRA, an MT system has to have the following two properties: (i) it has to be stratificational; (ii) it has to be multilingual.

(i) the translation relation must not hold between texts but should be split up into several simple translation relations holding between linguistically motivated representations: T1 . . . R1 . . . Rn . . . T2 (where each Ri is an artificial representational language and has to be an element of the set in (1)).

(1) Ri = {EMS, ECS, ERS, IS}
(EMS = EUROTRA Morphological Structure, ECS = Constitutent Structure, ERS = Relational Structure, IS = Interface Structure)

试读结束，需要全本PDF请购买 www.ertongbook.com