[*Translating and the Computer 8*. Proceedings of a conference… 13-14 November 1986, ed. Catriona Picken (London: Aslib, 1987)]

Research on machine translation at the University of Saarbrücken

Karl-Heinz Freigang

University of Saarbrücken, Federal Republic of Germany

INTRODUCTION

Research in the field of automatic analysis of language and machine translation has a long tradition at the University of Saarbrücken. In the late 1950s, a first attempt was made at the Institute of Applied Mathematics to develop a system for the automatic translation of Latin sentences (taken from a secondary school textbook) into German. In the early 1960s, a small group of researchers and students at the Institute of Applied Mathematics and the Institute of German Language and Literature, headed by Professor Hans Eggers, began to develop algorithms for the automatic syntactic analysis of a corpus of German texts, taken from newspapers and scientific textbooks (the RDE/FAZ-corpus). In the late 1960s and the early 1970s, this research group was asked by the Deutsche Forschungsgemeinschaft (DFG) to develop an automatic translation system from Russian into German on the basis of a Russian-English version of SYSTRAN. The idea of adapting this SYSTRAN version to German as a new target language was soon abandoned, and it was decided to develop an independent Russian-German MT system. This led to the foundation in 1972 of the ‘Sonderforschungsbereich 100, Elektronische Sprachforschung’ with the aim of developing the ‘Saarbrücker ÜbersetzungsSYstem’ (SUSY). In the following years, the Russian-German version of SUSY was taken as a basis for the integration of further language pairs, first French-German and in 1978 the English-German component. There were also attempts made at adapting SUSY to translation from Esperanto into German, plus German into English and French and to implement prototypes for the language pairs Danish-German and Dutch-German. The languages which, in principle, can be treated by SUSY are shown in Figure 1.

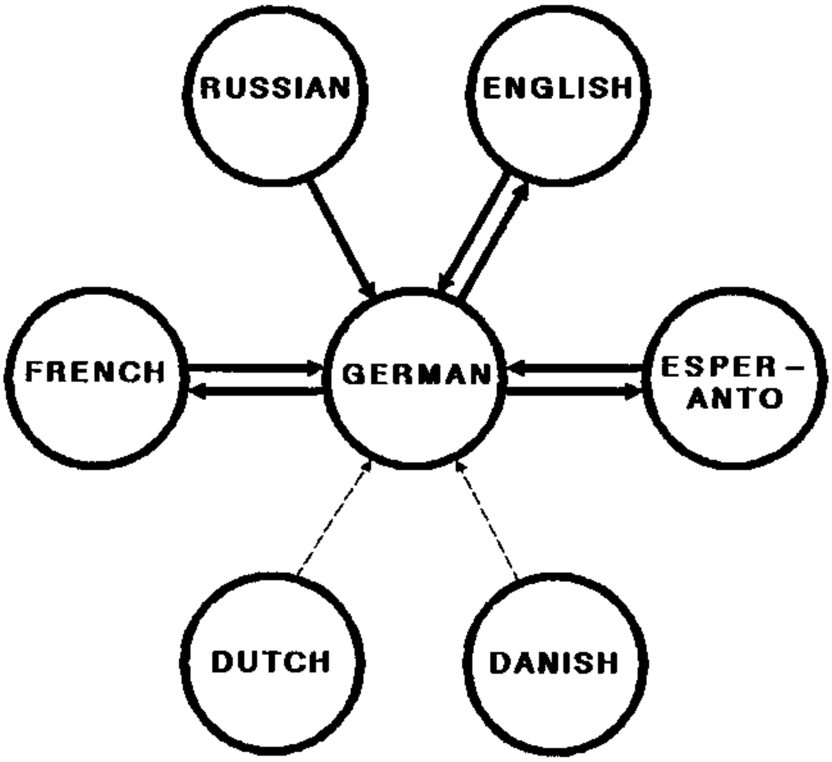

Figure 1. Languages and language pairs

THE SUSY MACHINE TRANSLATION SYSTEM

The theoretical linguistic basis of the SUSY machine translation system is a so-called base grammar, the underlying linguistic theory of which is a valency grammar with some additional elements of transformational grammar. This means that as the result of syntactic analysis each sentence of the source language text is assigned a ‘base structure’ (‘deep structure’), in which the verb (predicate) is represented as the central element with all other constituents of the sentence being described in their relation to the predicate. The internal structures of the constituents, i.e. the verbal and nominal groups, are also described as dependency (or valency) relations.

The SUSY machine translation system is conceived as a multilingual system based on a three-stage model of the translation process. Thus, the translation process is divided into three phases: analysis, transfer and generation (synthesis). During the analysis phase source language sentences are analysed exclusively on the basis of the grammar and dictionaries of the source language without taking into account the target language into which they are to be translated. The resulting structural description of the sentences is represented as a tree structure containing the syntactic and semantic information relevant for their further treatment in transfer and generation. The transfer phase consists mainly of a substitution of source language lexical items by target language equivalents with only some minor structural changes. The tree structure representation, with target language lexical items as terminal nodes, is then taken as the input into target language generation, which operates on this tree structure without referring to the source language. This approach guarantees the multilingual applicability of the system, with common analysis and generation components for all source and target languages and, of course, transfer components for each language pair. In the analysis phase linguistic rules are in some cases separated from the algorithms responsible for their application, so that new rules can be integrated into the system without the necessity of modifying the algorithms; in other cases the linguistic rules are integrated in the algorithms and are then activated by specific parameters, which are not bound to particular languages but to certain linguistic features (linguistic characteristics) shared by several languages. In principle, the same is true for generation, i.e. there is also one common generation component for all target languages. Transfer components are implemented for each language pair based mainly on bilingual transfer dictionaries.

Both analysis and generation are divided into strictly ordered modules, which are activated sequentially. This strict division into modules (the so-called ‘operators’) is to a large extent linguistically motivated, but it is also due to historical reasons, because the limited capacity of the main frame on which the system was first implemented did not allow the whole system to be loaded at once. Each of these modules has strictly defined input/output conditions, i.e. there are strictly defined intermediate structures which have to be delivered from one module to the following one. This means that modules can only be activated in a pre-defined order and that backtracking between the modules is not possible.

In the following, I will give a short summary of the modules which form the analysis, transfer and generation components of the system. The overall system design is outlined in Figure 2 (for a more detailed description cf. Blatt, Freigang, Schmitz, Thome, 1985).

The translation process

(a) *Analysis phase*

Source language analysis is subdivided into eight modules (‘operators’), which are activated sequentially.

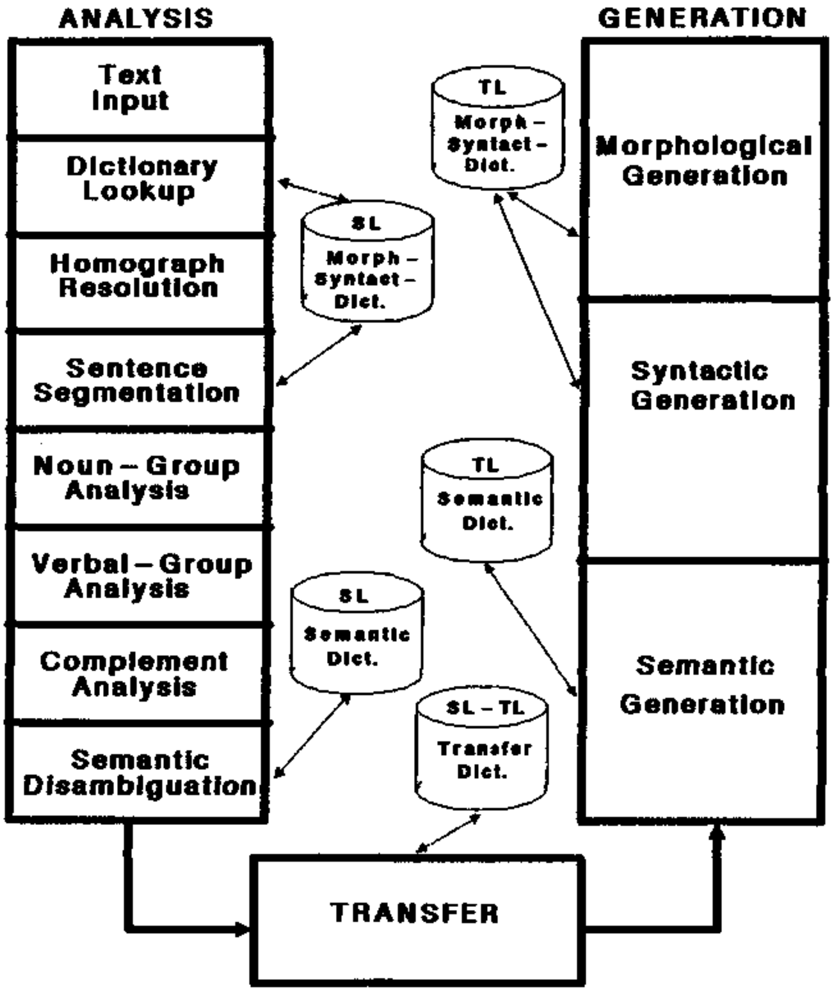

Figure 2. Outline of overall system design

Text input (operator LESEN). The task of this module consists of reading the source language text, splitting it up into a sequence of sentences on the basis of punctuation and further subdividing the sentences into sequences of isolated words, i.e. strings of characters separated from each other by blanks or punctuation marks.

Dictionary look-up and morphological analysis (operator WOBUSU). The aim of morphological analysis and dictionary look-up is to assign to each text word its possible syntactic readings (parts of speech, word classes) and the morpho-syntactic information necessary for syntactic analysis. The module makes use of a monolingual source language dictionary containing stem forms as well as word forms for not-inflected or irregularly inflected words. The morphological analysis algorithm separates stems from inflectional endings, identifies derivations consisting of stems and prefixes and/or suffixes, and recognises compounds consisting of two or more independent stems. The result of this step is a data structure which contains for each text word information records representing its word class readings and its syntactic information, coded in the dictionary (e.g. valency frames, types of complementation, etc.) (cf. Freigang, Schmitz, 1979a).

Homograph resolution (operator DIHOM). In the SUSY system, word forms belonging to more than one word class (as assigned during dictionary look-up) are called homographs. The word classes assigned to each word form of a sentence are the basis for the definition of larger syntactic units and thus for the identification of the syntactic reading(s) of a given sentence. In theory, the syntactic analysis could operate on as many strings of word classes as could be formed out of the results of dictionary look-up. In order to reduce the number of such strings of word classes as early as at the beginning of syntactic analysis, an independent module for the disambiguation of homographs has been introduced into the system. The task of this module is to resolve word class ambiguities by means of three different procedures. First, there are special routines for the resolution of certain types of ambiguities with high frequency (e.g. the ‘ing’ form in English). Secondly, there are the so-called (language-specific) ‘inhibition matrices’, which contain sequences of word classes not allowed in the language (e.g. definite article + finite verb). The third step makes use of tables of probability values weighting the frequency or probability of the occurrence of certain pairs of word classes. As the result of this procedure, up to 12 strings of word classes can be computed for each sentence, which are ordered according to their probability value and each of which in itself is unambiguous (cf. Freigang, Schmitz, 1979b).

Sentence segmentation (operators SEGMENT/PHRASEG). This module gets as its input the strings of word classes computed during the preceding step and operates on each of these strings separately with the aim of recognising clause boundaries within the sentence and identifying the types of clauses. This task is carried through by two different procedures depending on the language which is analysed. For the analysis of languages like German and Russian, where sentence segmentation is to a large extent dependent on punctuation, the module (operator) used is SEGMENT, based on the fact that in these languages clause boundaries are always marked explicitly by punctuation marks and/or conjunctions (cf. Luckhardt, 1980). For languages like English a different strategy had to be developed (PHRASEG), which is not based on punctuation but on linguistic features. The underlying assumption is that in a sequence of word classes containing two or more possible predicates there must be a clause boundary somewhere between them. Thus, for the identification of clause boundaries in English sentences a restricted syntactic analysis is necessary including a preliminary analysis of nominal and verbal groups as well as a preliminary valency analysis. The result of this module is a data structure containing for each reading of a sentence the clause boundaries, the types of clauses it consists of (e.g. adverbial clause, relative clause, object clause etc.) and the relations between the clauses (cf. Schmitz, 1986a).

Analysis of noun groups (operator NOMA). The task of this module consists of identifying the nominal elements within clauses, grouping them together into noun groups and describing the internal structure of these groups. The result is a data structure containing the noun groups together with a description of their internal structure (e.g. relations within complex noun groups, information on definiteness, number, gender, etc.) (cf. Blatt, 1983).

Analysis of verbal groups (operator VERA). The task of this step is to identify the verbal elements within the clauses, to combine them into verbal groups and to describe the internal structure of these groups. Verbal elements are for example: finite verb, infinitive, gerund, auxiliaries, modal verbs, etc. The result of this module is a structural description of all verbal groups occurring in the sentence including information on tense, mood, etc., attached to the governor of the group, i.e. the main verb, so that the lexical forms of auxiliaries are eliminated from the data structure, and only their morpho-syntactic information is maintained (cf. Freigang, 1986a).

Analysis of complements (operator KOMA). This module is the last step of the syntactic analysis. Its task is the description of the syntactic structure of the whole sentence, which consists of the description of the relations between the noun groups and the verbal groups either as complement relations (e.g. subject, direct object, prepositional object) or as adverbial relations. Subordinate clauses are related to their governing predicate with specification of valency relations (case relations or adverbial relations) or to their governing noun groups as attributive clauses. In this step some elements of transformational grammar are applied, such as transformation of passive constructions into their underlying active forms, in order to reconstruct the deep subject or deep object, which might be relevant for the subsequent semantic disambiguation; further transformations concern, e.g. the reconstruction of deleted subjects in infinite constructions. The result of this module is a complete description of the syntactic structure of the sentence, which can be represented as a valency tree reflecting the dependency relations between and within the sentence constituents (cf. Freigang, 1986a).

Semantic disambiguation (operator SEDAM). The task of this module consists of reducing the remaining syntactic ambiguities by making use of semantic restrictions, of resolving lexical ambiguities, of transforming prepositions which are not valency bound into an interlingua and of identifying idiomatic expressions. These tasks are carried through by means of a language-independent algorithm and language-specific semantic dictionaries. The dictionaries contain semantic features and a syntactico-semantic classification of nouns and disambiguation rules, which are interpreted by the algorithm. This module is the last step of the analysis phase of SUSY. Provided that all steps of the analysis have been applied successfully, the result is an unambiguous syntactic and semantic description of the sentence (cf. Blatt, 1981).

(b) *Transfer (Operator TRANSFER)*

Source language analysis is now followed by the transfer phase. Its task consists of the substitution of source language lexical items by target language equivalents by means of a bilingual transfer dictionary and of some minor modifications in structure caused by certain lexical items (e.g. changing direct object into prepositional object or vice versa). The result of this phase is the tree structure obtained during analysis with target language lexical items as terminal elements (cf. Luckhardt, Maas, 1983).

(c) *Generation*

During the generation phase (synthesis) target language sentences are generated on the basis of the tree structure taken over from transfer. It is subdivided into three modules (cf. Luckhardt, Maas, 1983).

Semantic generation (operator SEMSYN). While during semantic analysis source language prepositions have been transformed into expressions of a semantic interlingua, the main task of semantic generation consists of transforming these interlingua expressions into target language prepositions based on features of the nouns they are governing. The module makes use of a monolingual semantic dictionary of the target language, which contains the same types of entries as the source language semantic dictionary used in analysis.

Syntactic generation (operator SYNSYN). The syntactic generation module takes over the output of the semantic generation, in order to generate the syntactic surface structure of the target language sentence from the deep structure. It makes use of a monolingual dictionary of the source language, which provides syntactic and morphological information on target language words. The result is a linearised structure which is ordered according to the grammatical rules of the target language.

Morphological generation (operator MORSYN). The last step of the whole translation process is the module of morphological generation. Its task consists mainly of generating the inflected word forms of the target language sentence, according to the morphological information provided by the target language dictionary. Source language and target language texts are stored in two columns side-by-side, the source text on the left hand side, the target text on the right. The result can be displayed on screen or printed on paper.

Applications of SUSY at the University of Saarbrücken

During the last few years efforts have been made at the University of Saarbrücken to test the possibilities of the application of the research system in other projects or for other purposes.

The analysis component of SUSY, called SATAN (Saarbrücker Automatische Text-ANalyse), is used for the purpose of automatic indexing in the CTX system, developed at the Department of Information Science (cf. Zimmermann, Kroupa, Keil, 1983).

A further project, which was sponsored by the Federal Minister of Science and Technology from January 1982 till December 1983, has been carried through in cooperation with the Federal Language Service (Bundessprachenamt) with the aim of developing a modified version of the English-German component of SUSY for the translation of texts with restricted vocabulary and restricted syntax (maintenance requirement cards made available by the Bundessprachenamt) (SUSY-BSA) (cf. Keil, Wilms, 1985).

A follow-up project of SUSY-BSA was the project SUSANNAH (SUSY ANwenderNAH) (January 1984-June 1986) with the aim of testing the possibilities of incorporating a large amount of terminological data from external data banks into the SUSY translation process and of embedding the whole system into a user-oriented environment (cf. Wilms 1985).

Within the project MARIS (Multilinguale Aspekte von Referenz-Informations-Systemen = Multilingual Aspects of Reference-Information-Systems; Department of Information Science and Institute of Applied Information Research, IAI) a system is being developed (STS — Saarbrücken Translation Service), which aims at a three-level approach concerning the translation of reference information (e.g. titles of scientific publications): human translation, using and creating terminological databases, machine-aided human translation, using the dictionary look-up routines of SUSY for the creation of terminological databases, machine translation with SUSY incorporating word processing for post-editing purposes (cf. Kroupa 1986).

The German-English and English-German components of SUSY are applied in a project at the Institute of Applied Information Research (IAI) which is concerned with the automatic translation of titles of scientific papers from Japanese into German and vice versa with English functioning as ‘switching language’.

Some concluding remarks

It has to be pointed out once more that SUSY is still regarded as a research system. The main tasks of our research team during the last few years have consisted of testing and improving the linguistic performance of the system, especially as far as its multilingual applicability is concerned. This research work has led to some general conclusions, which I would like to summarise as follows:

— The overall design of the SUSY system permits relatively easy modifications, if errors can be clearly ascribed to certain modules, if additional rules have to be introduced for specific subtasks or new languages, or even if completely new strategies have to be integrated for clearly defined linguistic phenomena (as was the case, e.g. with sentence segmentation for English). This is due to the subdivision of the system into well-defined modules with strictly defined input/ output conditions.

— A major drawback of the system, however, is to be seen in the treatment of ambiguities in SUSY. Structural ambiguities are not represented in one data structure; this means that in the case of ambiguities at any stage of the analysis, the system produces as many complete structural descriptions of the sentence as there are ambiguities, and the subsequent modules have to treat these different readings of one sentence separately. In order to restrict the amount of time needed for the analysis of a text, the number of readings of one sentence transmitted from one module to the subsequent one must be restricted. This might in some cases result in the situation that the correct reading of a sentence cannot be preserved for subsequent analysis stages, not because the system could not analyse the sentence correctly but simply because of these technical restrictions.

— Last but not least, it is a well-known phenomenon that in the field of computational linguistics a certain stage of development is achieved in a rather short time at the beginning of the research work; further improvements of the system, however, concerning its linguistic performance take a lot of time and are often only to be achieved by means of a complete re-design of the system. Such a re-design of the SUSY system has been attempted in the experimental system SUSY-II, which tries to avoid the drawbacks mentioned above concerning the treatment of ambiguities by using charts as a data structure, which permits the parallel representation and handling of ambiguities. On the basis of SUSY-II a new software system for natural language processing is being developed (SAFRAN — Software and Formalism for the Representation and Analysis of Natural Language) (cf. Licher, Luckhardt, Thiel, 1986).

PERSPECTIVES FOR FURTHER RESEARCH

It seems to be a generally accepted point of view in the field of machine translation research that ‘fully automatic high-quality translation’ of random texts cannot be achieved in the near future. On the other hand, it is a generally accepted fact, too, that in translation practice the importance of machine-aided tools is steadily increasing. For those engaged in research, this means that they have to take into account not only the improvement of the linguistic performance of MT systems but also the improvement of the environment of such systems and the everyday working conditions of translators. Therefore, our research team at the University of Saarbrücken has decided to concentrate future research work on the following topics:

— Empirical investigations concerning the everyday work of translators in different types of institutions (freelance translators, independent translation services, translation services in industry and public organisations). As a result of these investigations, the requirements concerning the environment and equipment of translators’ work stations for different types of users will be defined. The results of these empirical investigations will be represented in a general theoretical model of man/machine interaction in the field of translation.

— This model of man/machine interaction, including the different tools needed by different types of users, is regarded as the basis for the development of an integrated translator’s work station. This means that the work station has to be designed in a way which allows the user to decide for himself or herself which tools are necessary for his or her kind of application. Tools to be included in the work station are for example: multilingual word processing, integration of terminological databases into the word processing system and/or the machine translation or machine-aided translation system, and a user-oriented maintenance system for machine translation dictionaries. Up to now small prototypes for word processing including split screen technique, term bank integration and dictionary maintenance system have been implemented (cf. Schmitz, 1986b).

— Technological change in the field of translation must be reflected in the training of translators and interpreters. A concept for the integration of computational linguistics, machine translation and linguistic data processing into the regular curriculum for translators and interpreters is being developed (cf. Freigang, 1986b). In the last few years courses have been carried through at the Department for Applied Linguistics, Translation and Interpreting dealing with subjects such as ‘Introduction to machine translation and machine-aided translation systems’ and ‘Introduction to word processing systems with practical exercises’. By further developing this component of the curriculum, we hope to give our students and future translators as solid a basis for their profession as possible.

REFERENCES

Blatt, A. (1981), *DOKUMENTATION K6:Semantische Disambiguierung als Komponerite des Übersetzungssystems Englisch-Deutsch.* Saarbrücken.

Blatt, A. (1983), *DOKUMENTATION K7: Nominalanalyse als Komponente des Übersetzungssystems Englisch-Deutsch.* Saarbrücken.

Blatt, A., Freigang, K.-H., Schmitz, K.-D., Thome, G. (1985), *Computer und Übersetzen. Eine Einführung.* Hildesheim.

Freigang, K.-H. (1986a), *DOKUMENTATION K8: Verbal- und Komplementanalyse im Übersetzungssystem Englisch-Deutsch.* Saarbrücken.

Freigang, K.-H. (1986b), ‘Sprachdatenverarbeitung in der Übersetzer- und Dolmetscherausbildung. Ein Entwurf zur Aktualisierung des Studiengangs’ in *Lebende Sprachen,* November 1986.

Freigang, K.-H., Schmitz. K.-D. (1979a), *DOKUMENTATION K2: Wörterbucherstellung, Wörterbuchsuche und Flexionsanalyse im Übersetzungssystem Englisch-Deutsch.* Saarbrücken, 3rd updated version, 1982.

Freigang, K.-H., Schmitz, K.-D. (1979b), *DOKUMENTATION K3: Homographenanalyse als Komponente des Übersetzungssystems Englisch-Deutsch.* Saarbrücken, 4th updated version, 1983.

Keil, G.C., Wilms, F.-J.M. (1985), *Untersuchungen zur anwenderorientierten maschinellen Übersetzung natürlicher Sprachen auf der Grundlage des Saarbrücker Übersetzungssystems SUSY – Erfahrungsbericht und Systemdokumentation, Forschungsbericht ID 85-003, Universität des Saarlandes, Projekt SUSY-BSA.* Saarbrücken.

Kroupa, E. (1986), ‘Multilingual Aspects of Reference Information Systems’, in Gerhardt, T.C. (ed.), *Proceedings of IAI—MT 86. I. International conference on the state of the art in machine translation in America, Asia and Europe (20 - 22 August 1986, Dudweiler).* Saarbrücken, pp. 227-241.

Licher, V., Luckhardt, H.-D., Thiel, M. (1986), ‘Konzeption, computerlinguistische Grundlagen und Implementierung eines sprachverarbeitenden Systems’. Paper presented at the 3rd International Colloquium on Machine Translation, Saarbrücken, 1-3 September 1986 (to be published in: Wilss, W., Schmitz, K.-D. (eds.), *Maschinelle Übersetzung. Methoden und Werkzeuge. Akten des 3. Intemationalen Kolloquiums des Sonderforschungsbereichs 100 'Elektronische Sprachforschung', Saarbrücken, 1-3 September 1986.* Tübingen, 1987, pp. 113-153.

Luckhardt, H.-D. (1980), ‘Automatische Segmentierung von Sätzen’, in Eggers, H. (ed.), *Maschinelle Übersetzung, Lexikographie und Analyse. Akten des 2. Intemationalen Kolloquiums, Saarbrücken, November 1979, II. Die Workshops (= Linguistische Arbeiten, Neue Folge, Heft 3.2).* Saarbrücken, pp. 69-71.

Luckhardt, H.-D., Maas, H.-D. (1983), *SUSY-Handbuch für Transfer und Synthese. Die Erzeugung deutscher, englischer oder französischer Sätze aus SATAN -- Analyseergebnissen (= Linguistische Arbeiten, Neue Folge, Heft 7).* Saarbrücken.

Schmitz, K.-D. (1986a), *Automatische Segmentierung natürlichsprachiger Sätze.* Hildesheim.

Schmitz, K.-D. (1986b), ‘Mensch-Maschine-Schnittstelle am Übersetzerarbeitsplatz’. Paper presented at the 3rd International Colloquium on Machine Translation, Saarbrücken, 1-3 September 1986 (to be published in Wilss, W., Schmitz, K.-D. (eds.), *Maschinelle Übersetzung. Methoden und Werkzeuge. Akten des 3. Intemationalen Kolloquiums des Sonderforschungsbereichs 100 'Elektronische Sprachforschung', Saarbrücken 1-3 September 1986.* Tübingen, 1987, pp. 309-321.

Wilms, F.-J.M. (1985), ‘SUSANNAH: Ein praxisorientiertes maschinelles Übersetzungssystem’ in *Sprache und Datenverarbeitung,* 1, 1985, pp. 37-46.

Zimmermann, H.H., Kroupa, E., Keil, G.C. (1983), *CTX — Ein Verfahren zur Computerunterstützten Texterschließung, Forschungsbericht ID 83-006.* Universität des Saarlandes, Saarbrücken.

AUTHOR

Karl-Heinz Freigang, Sonderforschungsbereich 100, ‘Elektronische Sprachforschung’, Universität Saarbrücken, D-6600 Saarbrücken 11, Federal Republic of Germany.