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

Edith Kroupa

Multilingual Aspects of Reference Information Systems (MARIS)

Universität des Saarlandes
Fachrichtung Informationswissenschaft
Im Stadtwald
6600 Saarbrücken
Tel.: 0681 / 302-3549

Edith Kroupa

Multilingual Aspects of Reference Information Systems

The project MARIS (Multilingual Aspects of Reference Information Systems) aims to set up the technical and organizational fundamentals for a central translation service for the field of "specialized information". This involves the development of a three-phase procedure for the computer-aided generation of target-language terminology and for the translation of texts from referential databases. "SUSY", the Saarbrücken Translation System is used as machine translation system.
In each phase the result is a high-quality translation.
Fields of application are e.g.: Construction (IRB), Technical Rules (DITR), Social Sciences. Further fields of application are being prepared: Material Science, Mechanical Engineering, Electrical Engineering and Patents.

Short description

Name

MARIS (Multilingual Aspects of Reference Information Systems); research project
STS (Saarbrücken Translation Service); service
SUSY (Saarbrücken Translation System); machine translation system

Status

Research project financed by the Federal Minister of Research and Technology with services for the field of specialized information

Type

SUSY: Translation system consisting of analysis, transfer and generation; analysis output: dependency structure with interlingual labels; number of grammatical rules not known as they form part of the program; multilingual: German/English; English/German; German/French (in prep.); French/German (in prep.)
STS: Rough translation with post-edition and terminology generation

Costs

At the moment 2.50 DM per title; abstracts and full text are calculated per word

Dictionaries

Analysis:	German:	142,000
	German compounds:	153,000
	English:	10,600
Transfer:	Ger-Eng:	70,000
	Eng-Ger:	12,800
Generation:	German:	14,500
	English:	2,600
Semantics:	German:	75.600
	English:	5.300

Implementation

FORTRAN IV/Assembler on SIEMENS 7.570, operating system BS2000
FORTRAN 77/C on NIXDORF TARGON /35, operating system UNIX SYSTEM V

1. The project MARIS and the Saarbrücken Translation Service

The project MARIS (Multilingual Aspects of Reference Information Systems) is a research project at the University of the Saarland - Department of Information Science - financed by the Federal Minister for Research and Technology. The aim of the project is the development of a prototype for the technical and organizational infrastructure of a central translation service in the field of "Specialized Information".

The project is organized into two parts. The scientific part of the project is being realized at the Department of Information Science at the University of the Saarland. This includes the planning and development of the Saarbrücken Translation Service (STS) which will be applicable to any subject field.

From the beginning of the project, the translation of textual entities from databases into English (for the time being just titles and descriptors) runs parallel to the development of the translation concept. This practical part is being carried out at the "Institute of the Society for the Promotion of Applied Information Research Inc. at the University of the Saarland (IAI)".

In the framework of the STS, the software developments at the University of the Saarland ("Sonderforschungsbereich Elektronische Sprachforschung" and the Department of Information Science), such as the Informative Translation System ITS, Saarbrücken Translation System SUSY, Computer-aided Text Indexing System CTX are transferred to a dedicated computer (NIXDORF TARGON/35).

At the end of the project, a three-phase procedure should have been developed which supports the translation of texts from any specialized subject fields.

2. Starting position

There is a continuously increasing number of scientific publications in English which shows that English has become the international scientific language. This gives rise to two opposite development trends in the German specialized information market. On the one hand, providers of German specialized information are compelled to translate their information into English in order to increase their international acceptance and to raise their share of the market to an international level. The German-speaking user is confronted with the problem that although he usually has enough passive knowledge of the English language to understand English texts, he will only find the desired information in data stocks, if he has enough active knowledge (without taking into account that an information broker could be consulted).

This results in the setting up of two versions of the same databank: One for the German market and the other for the international market. In such cases, the "German" databank only forms part of the "international" data bank. Consequently this leads to a demand for (inexpensive) translations in the field of specialized information.

The data stocks (especially titles and abstracts) are machine readable. To some extent, translations from German into English and English into German (and other languages) have already been made; even the controlled vocabulary with which the indexing of contents is realized, is partly available in different languages. Thus machine translation aids become a true alternative.

In the meantime, machine translation aids have been developed which make efficient translation possible. These aids support text generation (translation, post-edition) of foreign languages and lead to greater consistency in the field of terminology. Their use, however, only pays if there is a minimum amount of translations.

In the field of specialized information the translation of descriptors and titles gives rise to some problems:

Context: Ambiguities can only be made disambiguous by the context. When translating isolated terms, and even when translating titles, the given context does not suffice for the detection, let alone for the clearing up of ambiguities. If however, there are records which have already been intellectually classified, the subject field assignments can be valued. They can also be used for the automatic classification of the concepts derived from the texts; thus a learning system for automatic disambiguation is created which is dependent on intellectual preparatory work.

Subject fields: Databases cover subjects which themselves cover a wide range of subject fields (one of the most extreme examples is the database of Technical Rules); this of course affects the scope of the vocabulary which is to be processed. The translation of the subject field "construction" has shown that a certain saturation regarding vocabulary can only be achieved when about 100,000 titles have been translated. That means, that for the use of machine translation, comprehensive preparatory work regarding terminology is necessary for each new "subject field".

The generation of terminology: New fields of technology produce new terms, which means that the terminologist or translator must do a great deal of research to find the equivalents in the foreign language.

Especially in the field of terminology, the advantages of the model of a central translation service become obvious.

During the first half of the project MARIS it became evident that some of the terminology from the different subject fields overlap considerably (e.g. Building Law, administrative regulations, execution of construction work, standards, etc.). On the whole a centre for machine-readable terminology will be a considerable help for the translation process.

3. Translation concepts

The Saarbrücken Translation Service STS (STS-I, STS-II, STS-III) - oriented at the field of specialized information - is being developed and tested. In any case the result is high-quality translation.

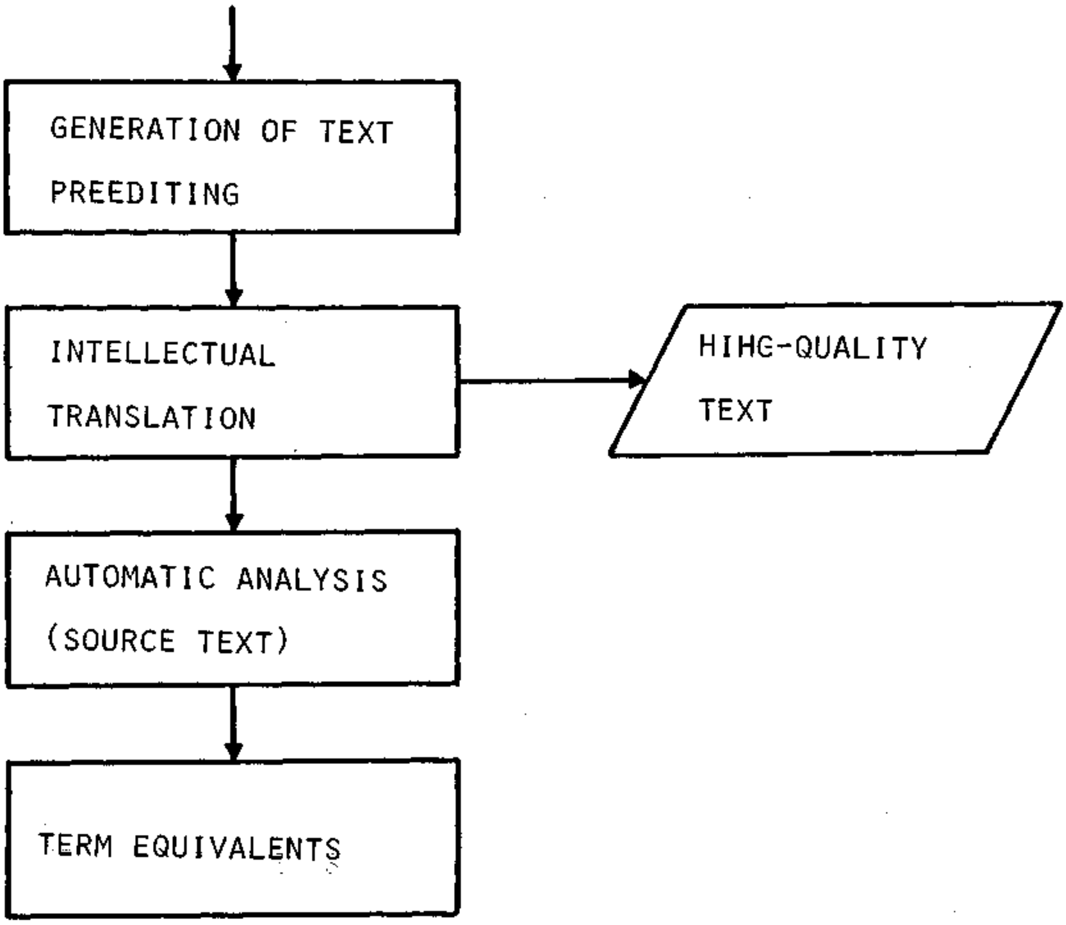
The efficient use of machine translation - excluding simple wordprocessing systems - is only possible if a minimum amount of machine-readable technical terminology is available. As this is not the case for the majority of the treated subject fields, the specific machine-readable terminology (for the source language as well as for the target language) must be generated before machine translation can be realized.

In phase two (STS-II) the intellectual translation is connected with automatic dictionary look-up. Phase three (STS-III) comprises postediting of rough machine translation for which automatic dictionary look-up and other wordprocessing tools are undertaken.

STS-III includes in any case the features of STS-II. The translation of texts which should not be translated by machine for structural, legal or other reasons, is at least to be provided with terminological equivalents for the source language terms.

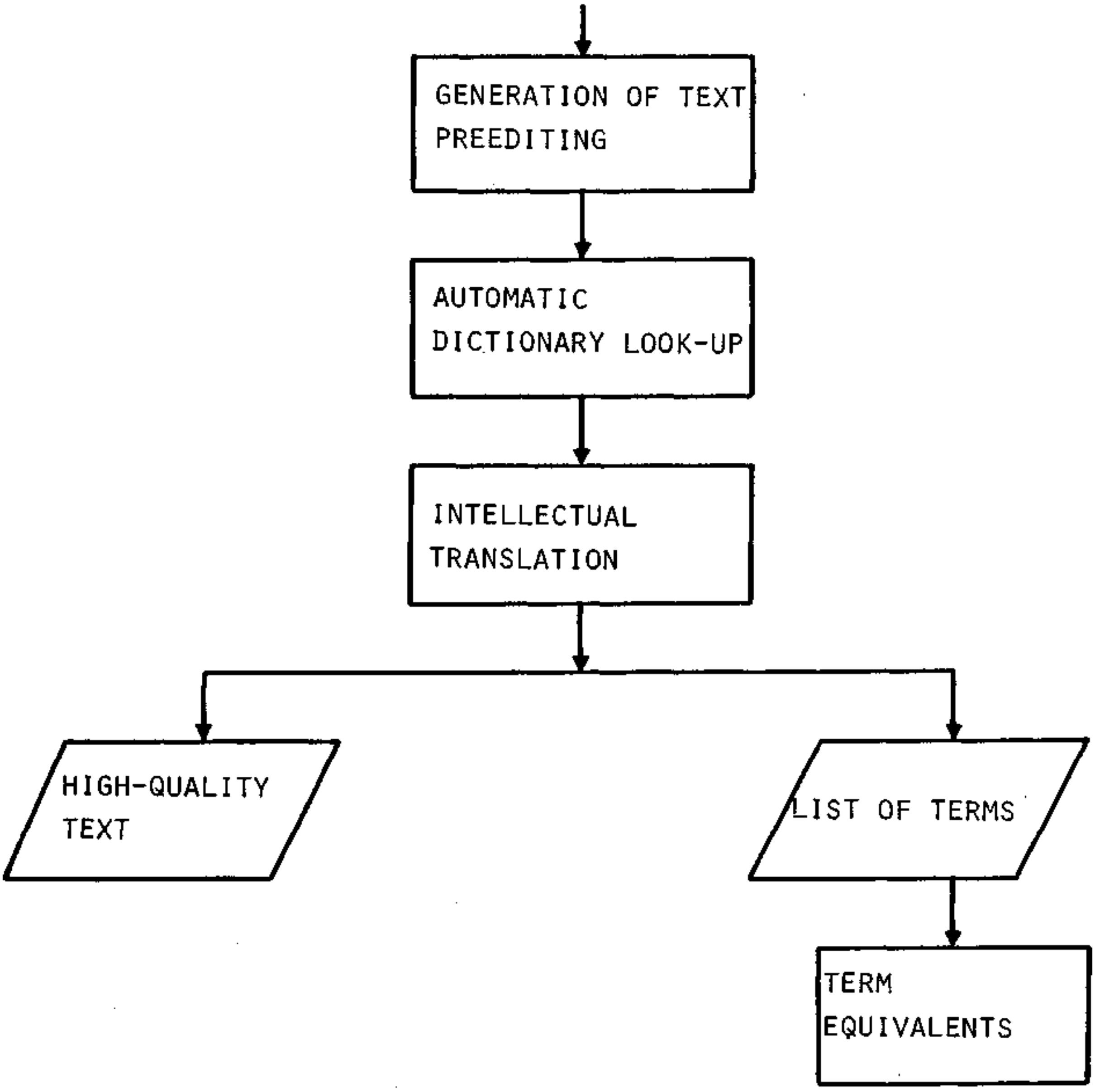
Intellectual translation with computer-aided terminology generation (STS-I)

The result of the first phase (STS-I) is a high-quality translation (HQT) with a simultaneous generation of machine-readable terminology. STS-I is being used for subject fields for which no sufficient machine-readable terminology is available. With the terminology derived from the translated texts, a database is set up which supports the retrospective generation of terminology. The terms (automatically reduced to their stems), together with their source-language and target-language contexts, are given to the translator. By means of interactive functions, the new terminology will be integrated in the automatic dictionaries with due consideration to each respective context.



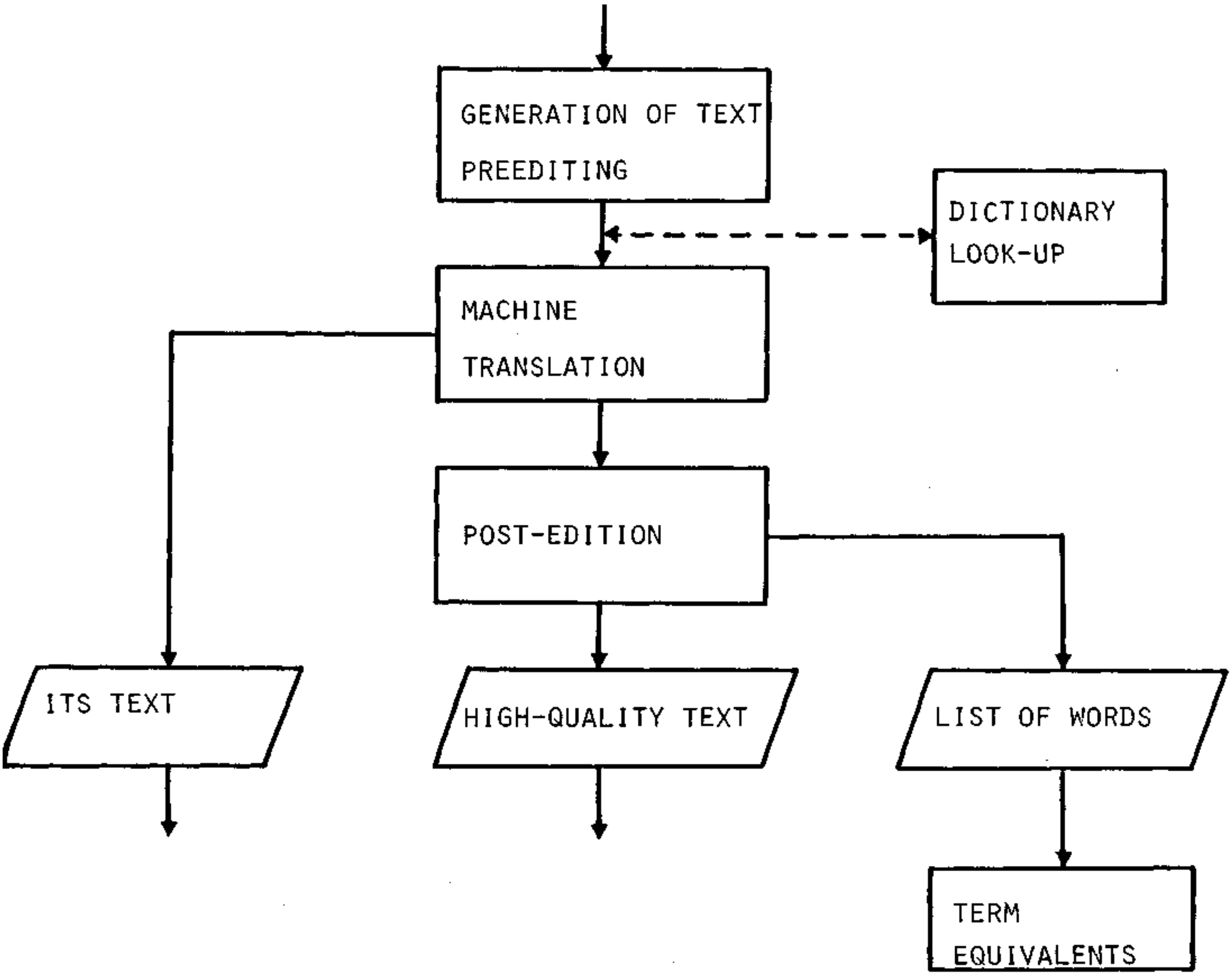
Intellectual translation with automatic dictionary-lookup (STS-II)

The result of STS-II is a high-quality translation as well. An efficient use of STS-II requires automatic dictionaries containing a sufficient amount of technical terms. In contrast to STS-I the translator is provided with the various equivalents of the terms of the text to be translated, given that the specific terms are contained in the automatic dictionaries. If no equivalent can be found to a certain term in the automatic dictionary, the term is listed with a special mark. Hence the translator is to translate the term and to add the new found equivalents with due regard to the context. The translation of texts which are not apt for rough translation can at least be supported by STS-II with as fas as terminology is concerned.



Machine translation with post-edition (STS-III)

The result of STS-III is a high-quality translation as well as an update of the automatic dictionaries (i.e. addition of new technical terms or equivalents). The basis of STS-III is rough machine translation (ITS-text). The work of the translator mainly consists of the post-edition of machine translated texts. Analogous to STS-II, the precondition for the realization of STS-III are almost "saturated" dictionaries which should require an addition of terms only in special cases. The Saarbrücken machine translation system "SUSY" is used for STS-III to its full extent. STS-III includes STS-II in so far as equivalents to the terms are given in the target language (even for texts which are not apt for rough translation).



4. Generation of terminology

An essential task in the framework of the practice-oriented development of a machine translation system, is the generation of dictionaries and specialized terminology, as well as their adaptation to the coding format required by the translation system. An exception is the German morphosyntactic dictionary of the Saarbrücken Translation System. This dictionary comprises about 140,000 entries and thus covers most of the German functional and basic vocabulary. The practical translation of texts of a subject field requires the expansion of the transfer and the synthesis dictionaries.

Supplementation by adopting existing machine-readable terminology

In exemplary investigations it became evident that due to the type of text to be translated (titles of new publications in the respective subject fields), a relatively large amount of terms must be expected which to date can neither be found in technical dictionaries, nor in terminology databanks. For example, up to now, there is no special terminology databank for construction; especially since the field of "construction" is difficult to define: ICONDA, the international construction database, for example, covers more than 15 subject fields. The field of "Technical Rules and Provisions", cannot really be described as one subject field either. The adaptation of terminology from other sources is not necessarily less expensive, not to mention the unresolved legal problems (copyright). The adoption of existing machine-readable terminology is therefore on the whole limited to the adoption of terminology stocks which were built up by the individual project partners. Up to now the following terminology stocks have been adopted:

o FINDEX Facet-oriented Indexing System for architecture and construction engineering from the Information Centre for Regional Planning and Building Construction (IRB) (about 6,800 terms);

- German/English controlled terms from the German Information Centre for Technical Rules (DITR) (about 11,000 terms).

Evaluation of translated text material

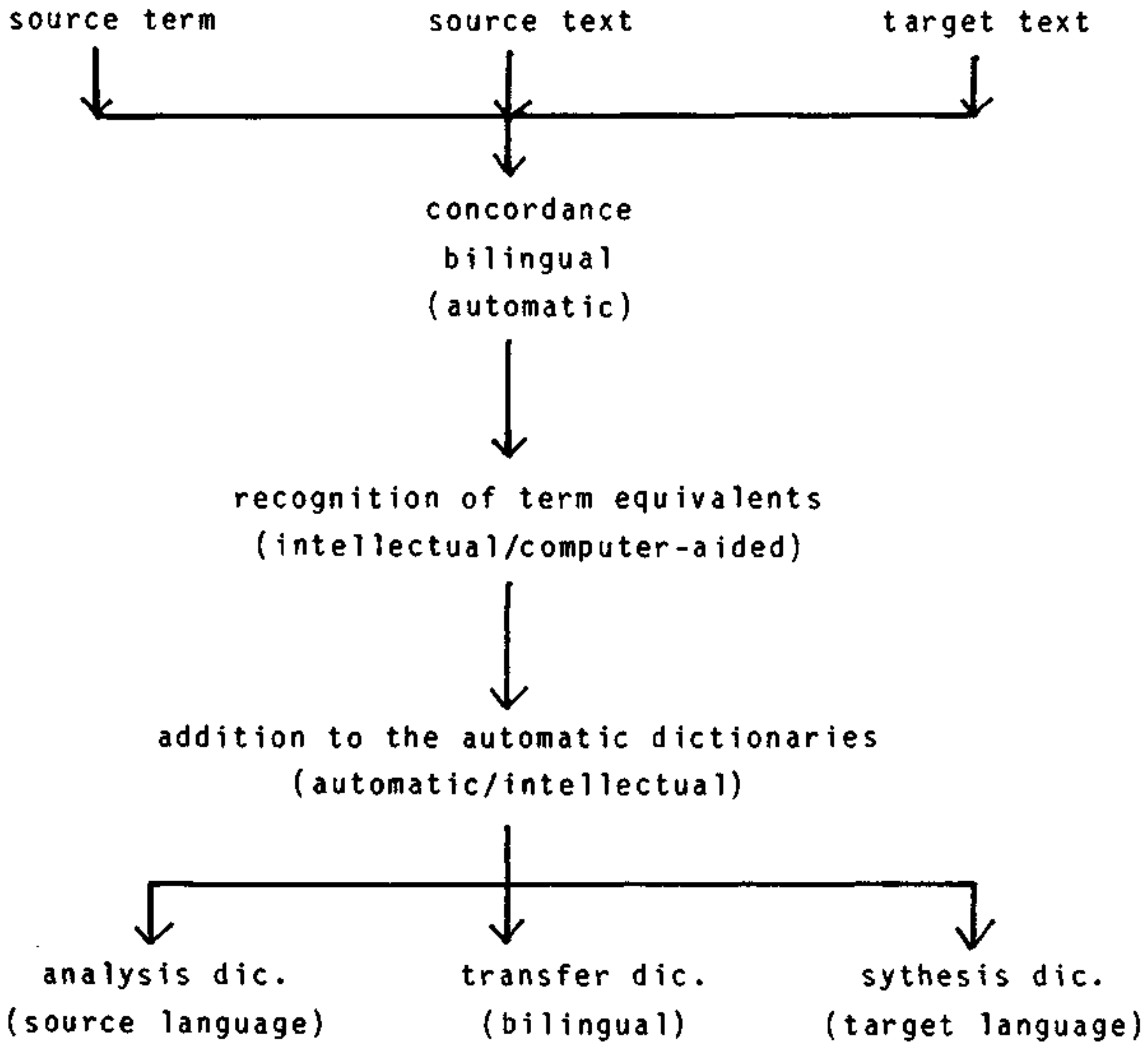
Bilingual text material is only of interest for the computer-aided generation of terminology if it is machine-readable. For the time being only bilingual data stocks which have already been translated by STS can be considered (for other machine-readable material similar conditions regarding the availability and the copyright of terminology apply). For reasons of personnel and time limits it has not yet been possible to edit multilingual machine-readable material which has not been translated by STS, although the process developed could be used for material from other sources as well.

Therefore terminology is being generated parallel to the translation work. The evaluation of the translated texts is realized by computer-aided text indexing of the original German text through CTX and by the setting up of a terminology databank. Every term is given a context in both languages, which is used by the translator to find the target language equivalent for the term in question. As a side effect of the selected procedure, the databank which is based on translated data can also be used as a terminology databank by the translators in the case of problems.

When each source language term has been translated, the translation must only be converted into the format of the system. If there are only bilingual equivalents of texts, a concordance is generated for the computer-aided allocation of the German term to the English equivalent.

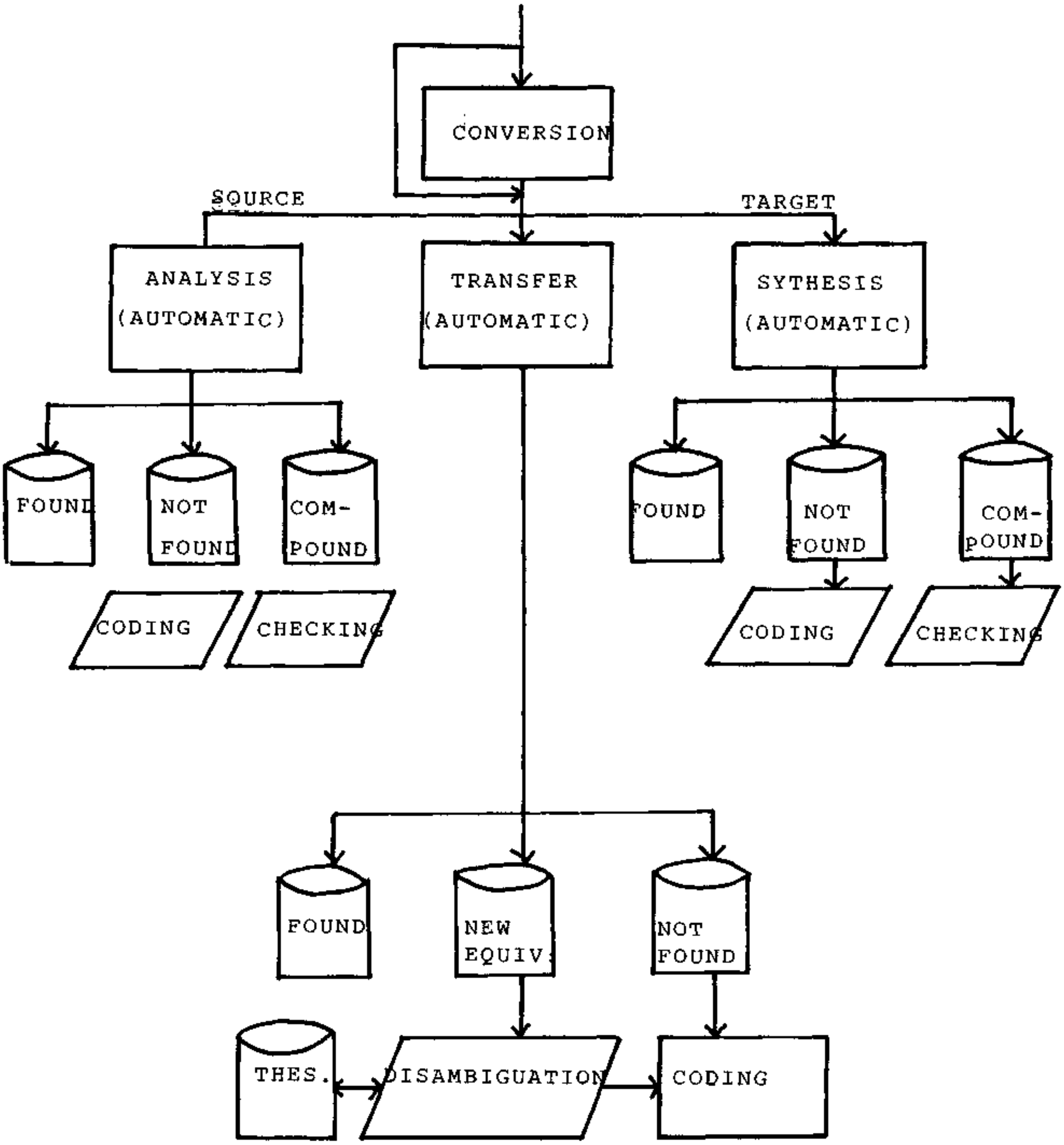
Derivation of equivalent terms from given bilingual texts

The German source terms (reduced to the stem) as well as the equivalents of texts in German and in the target language which have already been coordinated, are presented in machine readable form. The equivalents of the texts are allocated to the source terms (keyword out of context).



Coding of existing equivalents for the machine dictionaries

The source term and its target language equivalent are presented in machine-readable form. The updating of the automatic dictionaries consists of adding new terms and allocating new equivalents to existing entries.



References

Kroupa, Edith: CTX - Leistungsbeschreibung. In: Zimmermann, Harald H. (Hrsg.): Computergestützte Texterschließung mit CTX. Beiträge zum 1. Forum Informationswissenschaft und Praxis. Veröffent1ichungen der Fachrichtung Informationswissenschaft. Saarbrücken 1983.

Kroupa, Edith: Der Saarbrücker Translationsservice STS. Kurzfassung und Beispiele. Saarbrücken 1986.

Luckhardt, Heinz-Dirk; Maas, Heinz-Dieter: SUSY - Handbuch für TRANSFER und SYNTHESE. Linguistische Arbeiten des SFB. Neue Folge, Heft 7. Saarbrücken 1983.

Maas, Heinz-Dieter: Das Saarbrücker Übersetzungssystem SUSY. In: Sprache und Datenverarbeitung 1978 (1).

o.V.: MARIS - Mu1ti1inguale Anwendung von Referenz-Informations-Systemen. Projektbeschreibung. Saarbrücken 1985.

Peters, Jens-Peter: Übersicht über den Übersetzungsbedarf deutscher Fachinformationszentren. Saarbrücken 1986.

SFB 100, Projektbereich A (Hrsg.): SALEM - Ein Verfahren zur automatischen Lemmatisierung. Tübingen: Niemeyer 1980.